Visualization and Interactive Exploration of Spatio-Temporal and Thematic Information in Digital Text Archives

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

André Bruggmann

von

Degersheim SG

Promotionskommission

Prof. Dr. Sara Irina Fabrikant
(Vorsitz und Leitung der Dissertation)

Prof. Dr. Ross Stuart Purves

Dr. Katja Hürlimann

Zürich, 2017

Abstract

While rapidly growing unstructured and semi-structured online digital text archives (e.g., *Google Books*) potentially offer a wealth of useful and important information to all of us in the information society, limited access mechanisms hinder the effective and efficient extraction of interesting, meaningful, and relevant information from these data archives. Adopting a GIScience perspective in this thesis, we aim to provide interested information seekers with visual and interactive means to access relevant spatial, temporal, and thematic information, and latent structures found in large digital text archives, using a typical digital text archive in the humanities as a case study. Unstructured and semi-structured, now increasingly digitally accessible text archives from the humanities are particularly interesting for geographers, as they contain a wealth of spatial, temporal, and thematic information, largely untapped for spatio-temporal and thematic data analyses in geography to date.

We address this research challenge using a three-pronged approach, informed by state-of-the-art GIScience methods and techniques. First, we demonstrate that spatial (i.e., place names), temporal (i.e., dates), and thematic information (i.e., topics in text documents) can be automatically retrieved from the *Historical Dictionary of Switzerland* (HDS), as one typical, digitally available semi-structured text archive in the humanities. We then show that the retrieved information can be meaningfully transformed and reorganized using a spatialization approach, such that this information can be presented to information seekers in the humanities in two-dimensional spatialized displays for further data exploration. These spatialized displays visually uncover latent spatio-temporal and thematic structures in the HDS text archive. Finally, adopting a user-centered graphical interface design and evaluation approach, we integrate spatialized displays in interactive online web interfaces, to make reorganized spatio-temporal and thematic information from the HDS available to information seekers for further exploration and knowledge discovery.

For that we constructed *spatialized network maps* and a *spatialized thematic landscape map display* with spatio-temporal and thematic information automatically retrieved from the digital HDS. The *spatialized network maps* depict relationships between Swiss toponyms in different centuries based on how often toponyms co-occur in the same HDS articles. The *spatialized thematic landscape map display*, created based on the *self-organizing map* technique, displays HDS articles as points on a map where thematically similar articles

are placed closer to one another in the map than to semantically less similar articles. The maps can be explored interactively. To create useful and usable interactive web interfaces, including the spatialized displays, we involved target users early on in the interface design and development process. Target users provided valuable feedback in the performed *utility* and *usability* evaluations. This helped us to iteratively develop perceptually salient and cognitively supportive graphical user interfaces to the HDS text archive. It also facilitated access to and sense-making of the depicted information about the history of Switzerland.

This thesis has three major contributions: first, we provide a comprehensive text information retrieval approach going beyond existing approaches to extract information from text documents in the humanities and present a completely automatic approach to retrieve spatio-temporal and thematic information from a semi-structured text archive. Second, we illustrate how spatialization techniques can be used to depict spatio-temporal and thematic relationships and interconnections in the humanities, revealed by transforming and reorganizing the retrieved information. Third, we contribute a systematic user-centered method to incorporate the spatialized displays in interactive web interfaces. This allows interested information seekers in the humanities to explore spatio-temporal and thematic relationships and structures interactively, using advanced geovisual analytics approaches commonly known in GIScience, but still mostly unknown in history and the humanities.

The systematic evaluation of the automatically retrieved information from the HDS showed satisfactory quality, which suggests that this approach might be successful for other similar unstructured and semi-structured digital text archives in the humanities that include spatio-temporal and thematic information. Furthermore, the systematic evaluation of the constructed spatialized displays with target users suggests that using *spatialized network displays* to depict spatio-temporal relationships and interconnections, coupled with a *spatialized thematic landscape* to depict semantic similarities in text documents, aid target users in the humanities to gain new insights about spatio-temporal and thematic information buried in the HDS. The results of a final combined *utility* and *usability* study further reveals that target users are indeed able to interactively and visually explore the HDS text archive, and make sense of the novel spatialized displays.

In summary, this thesis highlights how advanced GIScience methods and approaches can be successfully transferred to the humanities to facilitate information access from growing unstructured and semi-structured text archives that also include spatio-temporal and thematic information.

Zusammenfassung

Einerseits stehen uns heutzutage immer mehr Informationen in unstrukturierten und semi-strukturierten digitalen Textarchiven (z.B. *Google Books*) online zur Verfügung. Andererseits fehlen uns oftmals effiziente Hilfsmittel, um interessante und sinnvolle Informationen aus diesen Textarchiven zu extrahieren. Ziel dieser Arbeit ist es, interessierten Personen einen visuellen und interaktiven Zugang zu räumlichen, zeitlichen und thematischen Informationen und versteckten Zusammenhängen in solchen Textarchiven zu ermöglichen. Dazu verwenden wir Methoden der *Geographischen Informationswissenschaften* (= GIScience), die wir auf ein typisches digitales Textarchiv in den Geistes- und Sozialwissenschaften anwenden. Für GeographInnen sind unstrukturierte und semi-strukturierte Textarchive der Geistes- und Sozialwissenschaften, die vermehrt digital verfügbar sind, besonders interessant, da sie eine Fülle von räumlichen, zeitlichen und thematischen Informationen beinhalten, die bisher nur sehr selten in räumlich-zeitlichen oder thematischen Studien in der Geographie analysiert wurden.

Wir stellen einen dreistufigen Ansatz vor, welcher auf Methoden und Techniken der GIScience zurückgreift. Zuerst zeigen wir, dass räumliche (z.B. Ortsnamen), zeitliche (z.B. Daten) und thematische Informationen (z.B. Themen von Textdokumenten) automatisch aus dem *Historischen Lexikon der Schweiz* (HLS), welches ein typisches digitales semi-strukturiertes Textarchiv der Geistes- und Sozialwissenschaften ist, extrahiert werden können. Dann zeigen wir, wie die extrahierten Daten transformiert und reorganisiert werden können, um zweidimensionale Darstellungen zu erstellen. Die Darstellungen basieren auf dem *Spatialization*-Ansatz und ermöglichen die visuelle Erkundung der in den Daten versteckten räumlich-zeitlichen und thematischen Zusammenhänge. In einem letzten Schritt wenden wir einen nutzerzentrierten Design- und Evaluationsansatz an, um die Darstellungen in interaktiven Webanwendungen interessierten Personen zur Verfügung zu stellen, und erleichtern es ihnen damit, neue Erkenntnisse zu räumlich-zeitlichen und thematischen Informationen und Zusammenhängen im HLS zu gewinnen.

Dazu haben wir *Netzwerk-Karten* und eine *Themenlandschaftskarte* erstellt. Die *Netzwerk-Karten* stellen Beziehungen zwischen Ortschaften der Schweiz in verschiedenen Jahrhunderten dar. Diese Beziehungen basieren auf der gemeinsamen Nennung der Ortschaften in Artikeln des HLS. Die *Themenlandschaftskarte* stellt HLS-Artikel als Punkte

auf einer Karte dar, wobei sich thematisch ähnliche Artikel näher beieinander befinden als thematisch unähnliche Artikel. Beide Karten können interaktiv erkundet werden. Um nützliche und nutzerfreundliche interaktive Webanwendungen dieser Karten zu erstellen, haben wir Personen unserer Zielgruppe früh in den Design- und Entwicklungsprozess der Anwendungen involviert. Die Personen haben uns wertvolle Rückmeldungen gegeben, die uns dabei geholfen haben, die Webanwendungen iterativ weiterzuentwickeln und perzeptuell sowie kognitiv ansprechend zu gestalten. Dies soll den interaktiven Zugang zu Informationen und mögliche Erkenntnisgewinne zur Geschichte der Schweiz erleichtern.

Diese Arbeit leistet drei wichtige Forschungsbeiträge: Wir stellen einen ganzheitlichen Ansatz vor, welcher bisherige Ansätze erweitert, indem er aufzeigt, wie räumlich-zeitliche und thematische Informationen komplett automatisch aus einem semi-strukturierten Textarchiv in den Geistes- und Sozialwissenschaften extrahiert werden können. Zusätzlich zeigen wir auf, wie der *Spatialization*-Ansatz zur Reorganisation der extrahierten Daten und zum Darstellen von räumlich-zeitlichen und thematischen Zusammenhängen in den Geistes- und Sozialwissenschaften genutzt werden kann. Ausserdem leisten wir einen Beitrag, indem wir einen systematischen und nutzerzentrierten Ansatz vorschlagen, der es erlaubt, die Darstellungen in interaktiven Webanwendungen Personen in den Geistes- und Sozialwissenschaften zur Verfügung zu stellen. Die dafür verwendeten Methoden sind in den GIScience verbreitet, haben jedoch bisher kaum Eingang in die Geschichtswissenschaften sowie allgemein in die Geistes- und Sozialwissenschaften gefunden.

Eine Evaluation hat aufgezeigt, dass die Resultate der automatischen Extraktion von räumlich-zeitlichen und thematischen Informationen befriedigend sind und dass der gezeigte Ansatz daher auf weitere unstrukturierte und semi-strukturierte digitale Textarchive in den Geistes- und Sozialwissenschaften, welche räumlich-zeitliche sowie thematische Informationen beinhalten, angewendet werden kann. Die systematische Evaluation der Karten hat gezeigt, dass die *Netzwerk-Karten* Personen dabei helfen, räumlich-zeitliche Zusammenhänge zu erkennen, und dass die *Themenlandschaftskarte* hilfreich ist, um thematische Ähnlichkeiten von Textdokumenten darzustellen. Eine Abschlussstudie zur Nützlichkeit und zur Benutzerfreundlichkeit hat ausserdem gezeigt, dass die Personen unserer Zielgruppe die interaktiven und visuellen Suchfunktionen in den Webanwendungen erfolgreich dazu benutzt haben, um neue Erkenntnisse über Raum, Zeit und Themen der Geschichte der Schweiz zu gewinnen.

Zusammenfassend kann gesagt werden, dass diese Arbeit darlegt, wie GIScience-Ansätze und -Methoden dazu genutzt werden können, um den Zugang zu unstrukturierten und semi-strukturierten Textarchiven in den Geistes- und Sozialwissenschaften, welche räumlich-zeitliche und thematische Informationen beinhalten, zu erleichtern.

Acknowledgements

This work was carried out during my time as a PhD student and research assistant at the Department of Geography at the University of Zurich. There are many people who supported me in completing this research project. I will name a few, but I would also like to express my deepest gratitude to everyone else that accompanied me on this path.

I am most grateful to Prof. Dr. Sara I. Fabrikant, who gave me the chance to work on this very interesting research project and supported me at all times with helpful feedback and advice. Prof. Dr. Sara I. Fabrikant supported me not only in research related matters, but also provided me with the opportunity to develop my teaching, computational, and transferable skills at the University of Zurich. I am very thankful for the trust, encouraging and motivating words, and inspiring discussions over the past years.

I also would like to express my sincere gratitude to the rest of my thesis committee, Prof. Dr. Ross S. Purves and Dr. Katja Hürlimann, for their continuous support and provision of valuable feedback on my work from very different scientific perspectives, as well as their insightful comments and encouragement.

A very special thank you is directed to Dr. Marco M. Salvini, who supervised my bachelor's and master's theses and motivated me to undertake a PhD. Furthermore, I would like to thank all members of the *GIScience Center* at the Department of Geography at the University of Zurich and particularly the *Geographic Information Visualization & Analysis* (GIVA) unit for the stimulating discussions and their motivating words. I would also like to mention some people separately to express my sincere gratitude: Dr. Ali Soleymani for encouraging and motivating me throughout my PhD and all the fun moments we shared outside of the work environment. Sara Maggi, my office mate, for the very pleasant and cheerful office atmosphere. Annina Brügger, Sascha Credé, Ismini Lokka, Benjamin Flück, Irene Johannsen, Dr. Stefano De Sabbata, Dr. Paul Crease, and Kenan Bektaş, for all the lunchtime and other discussions regarding research and anything else important in the life of a PhD student. Furthermore, I would like to thank Dr. Arzu Çöltekin, Dr. Kai-Florian Richter, Dr. Halldór Janetzko, Dr. Tumasch Reichenbacher, Dr. Curdin Derungs, Dr. Damien Palacio, Dr. Jannik Strötgen, and Julian Zell for all your valuable input to my research, and for all the inspiring and motivating discussions.

I owe a huge debt of gratitude to my entire family, particularly to my parents Renate Bruggmann and Guido Bruggmann and my brother Marc Bruggmann. Without their continuous support during my life in any aspect of my personal and academic path and all the encouraging words and love they gave me, I could not have reached this point in my life. Also, I would like to thank, from the bottom of my heart, Nadine Schwarz for her patience, understanding, and all of the support and encouragement she provided throughout my PhD. I also would like to express my gratitude to Ursula Schwarz, René Schwarz, and Philipp Schwarz for all the inspiring and motivating conversations and very enjoyable moments together. I would also like to thank my friends very much which were always there for me throughout my studies. I would like to mention some of them in particular that allowed me to experience countless happy and cheerful moments: Julian Lindenmann, Amir Habchi, Massimo Calamassi, Sarah Meier, Fabienne Forrer, Matthias Zumkehr, Oliver Deseö, Patrice Frei, Ramón Huldi, Michael Pichlmeier, Kaspar Fischer, Martin Zahner, and Simon Etter.

Lastly, I would like to thank all participants of the user studies for their very valuable feedback for my research.

Contents

Table of content

List of figures

List of tables

1 Introduction

1.1 Motivation

The Spanish sociologist Manuel Castells has defined the current time period as the information age, due to the rapid development of information technologies since the 1970s and the shift from an industrial to informational society in recent decades (Castells, 2010). As a result of this shift, the internet was commercialized which revolutionized our way of communicating (Castells, 2010: 45-53). People became increasingly connected within a global network. The *United Nations* specialized agency for information and communication technologies ITU (*International Telecommunication Union*) estimates, that in 2015, more than 40% of the world's population (about 3.2 billion people) was using the internet, compared to less than 10% in 2005 (ITU, 2015: 2). Technological advances and the spread of the internet resulted in an increase in the production of data volumes, which have been stored and shared online in recent years. Van den Bosch et al. (2016) estimate that more than 40 billion web pages were indexed by *Google* in January 2015, and Cisco (2015: 1) predicts that "annual global IP traffic will pass the zettabyte [= 1000^4 gigabytes] threshold by the end of 2016".

Domo's (2015) visualization (see Figure 1) depicts this wealth of information by illustrating various online services and how often they are used every minute of the day. This visualization does not only show the minutely data load of the online services, but also highlights the very different types of data which are produced. For example, social networks such as *Twitter*[1] or *Facebook*[2] generate large amounts of multimedia content (e.g., text, photo, video), Apple's *App Store*[3] provides access to apps (i.e., software), and *Netflix*[4] offers streaming video content. All services in Figure 1 involve internet users that create data by interacting with the service by, for example, liking posts on *Facebook*, sending *Tweets* on *Twitter*, or calling friends using *Skype*[5].

[1] Twitter: https://twitter.com/ (accessed April 2016)
[2] Facebook: https://www.facebook.com/ (accessed April 2016)
[3] App Store: http://www.apple.com/macos/what-is/ (accessed April 2016)
[4] Netflix: https://www.netflix.com/ (accessed April 2016)
[5] Skype: http://www.skype.com/ (accessed April 2016)

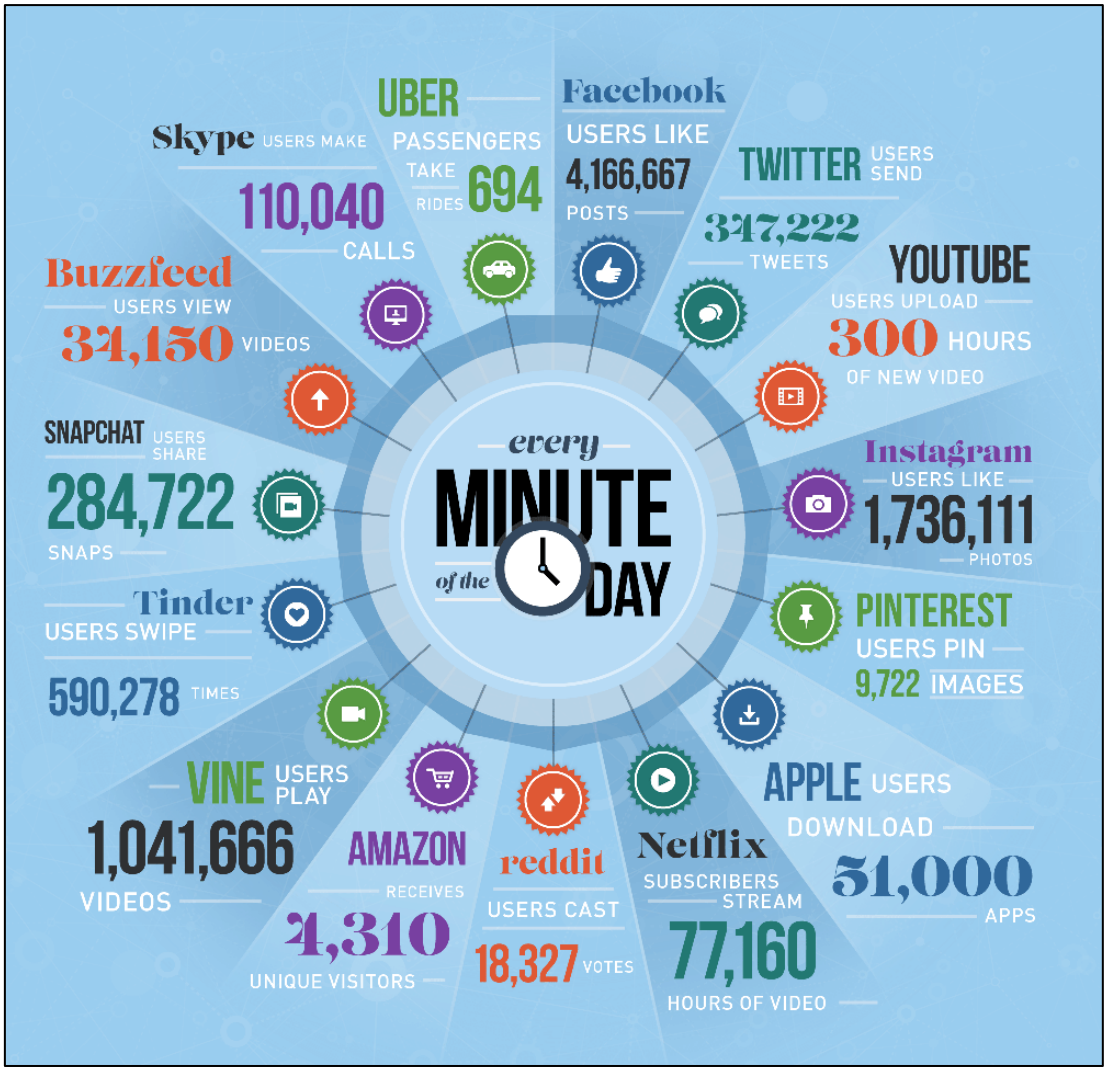


Figure 1: Amount of data produced by various online services (modified from Domo, 2015).

Open data initiatives are further important drivers of the online data growth. Hossain et al. (2016) situate open data initiatives in governmental, public (including private organizations), or mixed contexts, and identifies leading politicians, institutional pressures (e.g., pressure to release publicly-funded experiment data) as well as technological developments as main drivers for the initiatives. *OpenStreetMap*[6] and *Wikipedia*[7] are examples of open data projects at a global scale in a public context. *OpenStreetMap* aims at creating a free, editable map of the world, whereas *Wikipedia* is a free-access and free-content online encyclopedia. *Opendata.swiss*[8] is an example of a governmental open data initiative at a national scale. The *opendata.swiss* platform was launched in 2016 and provides free access to data from various political institutions in Switzerland.

[6] OpenStreetMap: https://www.openstreetmap.org/ (accessed April 2016)
[7] Wikipedia: https://www.wikipedia.org/ (accessed April 2016)
[8] opendata.swiss: https://opendata.swiss/ (accessed April 2016)

The large-scale digitization of information which has not been previously stored digitally is another driver of online data growth. Non-digital resources (e.g., books, images, maps, videos) are now digitized by (state) institutions, organizations, and companies (e.g., *Google*) and made available in massive online archives (Yang and Li, 2016). For example, *Google Books*[9], *The Universal Digital Library*[10], and *Project Gutenberg*[11] provide access to millions of digitized books, whereas the *Library of Congress*[12] and the *Digital Collections* of the *New York Public Library* (NYPL)[13] provide access to digitized books, photographs, maps, manuscripts and many more digitized resources.

As a consequence of the current wealth of online information due to online services, open data initiatives, large-scale digitization projects, and others, an interdisciplinary research community has been attracted to study and analyze these data. While librarians and economists are interested in societal and business model transformations arising from digitization and *big data* analytics (e.g., Loebbecke and Picot, 2015), (digital) humanities researchers produce lengthy discussions regarding the conflicting needs of the quantitative and qualitative methods used to analyze *big data* (e.g., Gooding, 2013). Other fields of research such as *data mining* and *knowledge discovery* search for new ways to structure information adequately in order to gain new insights about data and the interconnection of items within it (e.g., Fayyad et al., 1996).

For the humanities, large online text archives (e.g., *Google Books*) are particularly interesting, as texts have been central to certain humanities disciplines (e.g., history, language, and literature) long before the digitalization movement. Text archives in the humanities are also interesting for geography, and particularly for GIScience, since they often contain a great deal of spatial, temporal, and thematic information that remains largely untapped for spatio-temporal and thematic analyses. However, accessing information in these text archives is challenging because the texts are often *unstructured* (i.e., raw text) or *semi-structured* (i.e., raw text including tags to identify certain elements within the data). An example for an *unstructured* text archive in the humanities could be a history book collection about the evolution of a country which has been digitized using *optical character recognition* scanning software, stored online, and made available on a freely accessible web page. Digitized books are supposed to contain many *toponyms* (e.g., names of cities, regions, or countries), temporal references (e.g., dates), and cover many different themes (e.g., economy, politics, society, conflicts, and wars). However, spatial, temporal, and thematic information must first be made explicit before spatio-temporal and thematic analyses can be conducted. For example, *toponyms* must be identified in the books, then retrieved and stored in such a way that they can be queried. Only then may the occurrences of *toponyms* in books be further analyzed. Manually identifying and retrieving information from large text archives is too time consuming, and thus automatic methods are needed.

[9] Google Books: https://books.google.com/ (accessed April 2016)
[10] The Universal Digital Library: http://www.ulib.org/index.html (accessed April 2016)
[11] Project Gutenberg: https://www.gutenberg.org/ (accessed April 2016)
[12] Library of Congress: https://www.loc.gov/library/libarch-digital.html (accessed April 2016)
[13] Digital Collections of the NYPL: http://digitalcollections.nypl.org/ (accessed April 2016)

The challenge inherent in automatically identifying and structuring spatial, temporal, and thematic information in large unstructured and semi-structured online text archives in the humanities represents the starting point of this thesis. Based on this challenge, the following section introduces the problem statement of this thesis.

1.2 Problem statement

Franco Moretti, a well-known (digital) humanities and literary history scholar, builds upon the challenge of identifying and structuring information in large text archives in the introduction of his book *Graphs, Maps, Trees*.

> *"(…) literature, the old territory (more or less), unlike the drift towards other discourses so typical of recent years. But within that old territory, a new object of study: instead of concrete, individual works, a trio of artificial constructs—graphs, maps and trees—in which the reality of the text undergoes a process of deliberate reduction and abstraction. 'Distant reading', I have once called this type of approach; where distance is however not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.*
> *(…)*
> *And so, while recent literary theory was turning for inspiration towards French and German metaphysics, I kept thinking that there was actually much more to be learned from the natural and the social sciences."*
>
> Moretti (2005: 1-2)

Moretti (2005: 1-2) argues for the use of reduced information in texts (i.e., *fewer elements*) to depict them in abstracted views such as *graphs*, *maps*, and *trees*. These abstracted views of texts help humanities scholars to understand overall interconnections, shapes, relations, and structures in the text archives (Moretti, 2005: 1-2). Moretti (2005: 1) calls this concept *distant reading*. Similarly, Jockers (2013) investigates *distant reading* while further highlighting the importance of *close reading* (i.e., studying single text documents) to analyze large text archives in the humanities. Jockers (2013: 27) argues that "…by exploring the literary record writ large, we will better understand the context in which individual texts exist and thereby better understand those individual texts". Therefore, this refers to not only the identification and structuring of information in texts, as shown in the last section, but also the presentation of information through visual means (e.g., *graphs*, *maps*, and *trees*) and on different information levels (i.e., *distant* and *close reading*) being important to the humanities in order to gain new insights and generate hypotheses regarding text data.

In recent years, the *digital humanities* community has evolved. *Digital humanities* investigate the combination of humanities approaches using computational methods (Kaplan, 2015). Both Moretti (2005) and Jockers (2013) contributed to this community with their idea of *distant* and *close reading*. Furthermore, the *digital humanities* community debates the adequate presentation of information in text archives to potentially interested users in the humanities. The use of visual and dynamic user interfaces is one option. For

example, Drucker (2011b) discusses interfaces and relates humanities approaches to *interface theory*, while Kirschenbaum (2004) discusses *aesthetics* and the *usability* of interfaces in the humanities.

The importance of geographic (i.e., spatial, temporal, and thematic) information and concepts in order to study the humanities from a geographic perspective has been highlighted by several sub-disciplines of *digital humanities* at the nexus between geography and the humanities, which have evolved over recent years. For example, *GeoHumanities* seeks to connect the concepts of *space* and *place* with the humanities (e.g., Dear, 2015, Hawkins et al., 2015), and the s*patial humanities* introduces GIS technologies to the humanities (e.g., Bodenhammer et al., 2010, Bodenhammer et al., 2013). *Historical GIS* follows a similar approach to the *spatial humanities*, but with a specific focus on historical data. This is particularly interesting for spatio-temporal and thematic analyses, as many historical data sources and text archives contain a large amount of spatial, temporal, and thematic information (e.g., Gregory and Ell, 2007, Gregory and Healey, 2007).

Building upon the arguments in the previous and current section, we define the following problem statement.

Problem statement

Many large (online) archives have been generated in recent years, and, in the humanities, many contain unstructured or semi-structured text documents. Recent research endeavors in *digital humanities* illustrate the interest of the humanities in studying spatial, temporal, and thematic information, structures, and interconnections in text archives. However, automatically identifying and structuring spatial, temporal, and thematic information in these text archives and also providing adequate (visual) presentation of the information to interested information seekers in the humanities represents a challenge and the subject of ongoing research in *digital humanities*.

This thesis addresses this challenge and follows an approach which incorporates all steps, from raw data processing of a large text archive in the humanities to the presentation of spatial, temporal, and thematic information to interested information seekers in the humanities while considering both the *distant* and the *close reading* concept.

Research questions were defined based on this problem statement and are presented in the following section along with the research approach for this thesis.

1.3 Research questions and research approach

In *Section 1.1*, we illustrated the challenge of identifying and structuring spatial, temporal, and thematic information in unstructured and semi-structured text archives in the humanities, which motivates the first research question.

Research Question 1

How can information about space, time, and theme be automatically retrieved from unstructured and semi-structured text archives in the humanities so that hidden structures and relationships can be uncovered in the data?

In order to answer *Research Question 1*, we applied *geographic information retrieval* (GIR) methods in this thesis. GIR is considered relevant to address *Research Question 1* as it provides well-established approaches to automatically retrieve spatial, temporal, and thematic information from unstructured and semi-structured content, and may thus be applied to unstructured and semi-structured text archives in the humanities. Latest methodologies and GIR methods relevant to this research project are presented in *Section 2.1*.

Next, the retrieved spatial, temporal, and thematic information has to be analyzed, transformed, and reorganized in order to present it to interested information seekers in the humanities via visual displays which support the exploration of the information and the generation of hypotheses. As Moretti (2005: 1) points out, reduction and abstraction of texts does not only allow for the production of simple summaries of facts (e.g., frequency of occurrence of toponyms in texts), but also opens up the possibility to grasp and depict interconnections of spatio-temporal and thematic elements as well as implicit relations and structures of these elements in text data, as discussed in the previous section. This motivates the second research question.

Research Question 2

How can we spatialize uncovered spatio-temporal and thematic structures and interconnections extracted from unstructured and semi-structured text archives in the humanities?

In order to answer *Research Question 2*, the *spatialization framework* could be useful as it offers a systematic approach to transform high-dimensional numerical and non-numerical data into lower-dimensional, spatial visualizations using *spatial metaphors* (Skupin and Fabrikant, 2007). The *spatialization framework* is introduced in *Section 2.2* of this thesis.

The spatialized displays produced in response to *Research Question 2* should reach target users and facilitate them to gain new insights and generate hypotheses regarding the spatio-temporal and thematic information as well as structures and interconnections

present in the data (i.e., *distant reading*). In addition, target users should have the possibility to access the raw data source (i.e., *close reading*). This motivates the third research question.

Research Question 3

How can we make spatialized information about space, time, and theme from unstructured and semi-structured text archives available to information seekers in the humanities to support sense-making and the generation of new insights about these text archives?

In order to answer *Research Question 3*, *geovisual analytics* was identified as being relevant because this research field suggests methods for including spatialized displays in interactive and exploratory (web) interfaces, involving target users early on in the interface development process. Involving target users early on in the interface design and evaluation process is important in order to determine the requirements of the target users and how these requirements can be incorporated into web interfaces that support the information-seeking process of target users. *Geovisual analytics* is introduced in *Section 2.3* of this thesis.

In Figure 2, the overall research approach for this thesis is summarized. The first three stages of the approach are related to the research questions. The final stage considers the output of the research project (i.e., *geovisual analytics* interfaces) and aims to provide target users with access to exploratory web interfaces. These interfaces incorporate *distant* and *close reading* functionalities and should motivate target users in the (digital) humanities to explore large text archives, gain new insights, and develop new hypotheses regarding spatial, temporal, and thematic information and interconnections in unstructured or semi-structured digital text archives.

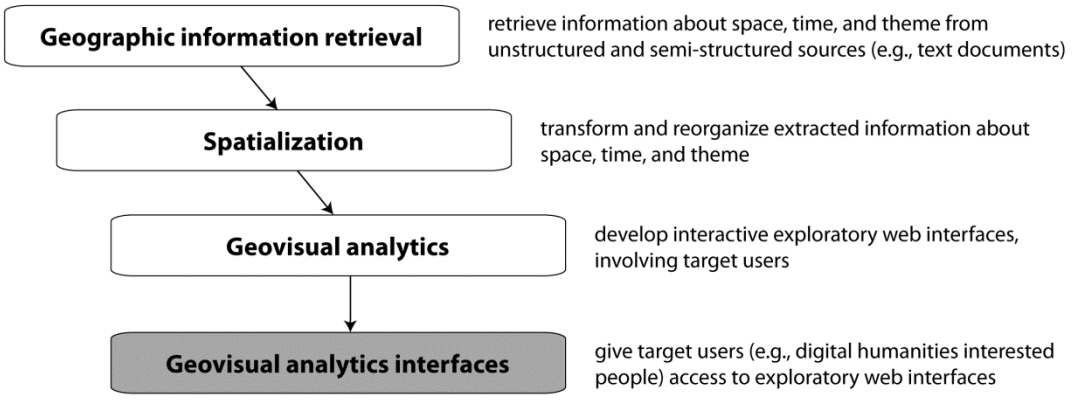


Figure 2: The stages of the PhD project and their contribution to this thesis (modified from Bruggmann and Fabrikant, 2016: 3).

The research approach outlined in Figure 2 illustrates the contribution of this thesis to *digital humanities*. In addition, we aim at contributing to the *GIScience* and particularly to the *geovisual analytics* community, as we provide a systematic approach which

incorporates all steps: from the automatic retrieval of spatial, temporal, and thematic data from unstructured and semi-structured text archives in the humanities, the transformation and visualization of the data according to the *spatialization framework*, and the development and empirical evaluation of exploratory and interactive web interfaces which involves target users in the interface design process. Furthermore, our research approach aims at contributing to the *geographic information search* field. Ballatore et al. (2015: 6) highlight that "spatial, temporal, and thematic information are always interacting in search(…)". Therefore, "(…) search should be able to integrate and combine these dimensions" (Ballatore et al., 2015: 12). In addition, Ballatore et al. (2015) highlight the need for further evaluation of exploratory search in the context of unstructured data and conclude that "…geography can benefit from new search approaches to explore data and formulate new research questions" (Ballatore et al., 2015: 15). Both the combined spatial, temporal, and thematic search as well as the evaluation of exploratory search is addressed in this thesis.

1.4. Structure of the thesis

The thesis is structured following the stages presented in Figure 2. In *Chapter 2*, we illustrate state of the art and related work relevant to this thesis in *geographic information retrieval* and *geovisual analytics*. Furthermore, we describe the *spatialization framework* which is applied to visualize the spatio-temporal and thematic data in spatialized displays in this project. In addition, we introduce the *digital humanities* field. In the final section of *Chapter 2*, the research gap, which was identified by analyzing current methodologies, is presented.

The *Historical Dictionary of Switzerland* (HDS)[14] is a typical large and semi-structured text archive in the humanities and contains a large amount of spatial, temporal, and thematic information. Therefore, the HDS fits well as a case study in this project. We introduce the HDS in *Chapter 3*. The methodological framework applied to the HDS is presented in *Chapter 4*. We highlight the retrieval of spatial, temporal, and thematic data from the HDS, illustrate the visualization of this data in spatialized displays, and present a user-centered design and evaluation approach to develop exploratory and interactive web interfaces. In *Chapter 5*, we detail the characteristics of the spatial, temporal, and thematic data retrieved from the HDS. Secondly, we present the spatialized displays, and thirdly, we describe the prototype web interfaces developed in this project. We also highlight the importance of involving target users' feedback in the interface design process and present the results of a combined *utility* and *usability* study to evaluate prototype implementations. In *Chapter 6*, we evaluate the *quality* and the *sensitivity* of our results to the methods we chose and parameters we applied. In *Chapter 7*, we discuss the results of this project and relate our findings to previous work in fields relevant to this project. In *Chapter 8*, we outline the main achievements of this thesis and identify areas of potential future work.

[14] Homepage of the Historical Dictionary of Switzerland (German version): http://www.hls-dhs-dss.ch/d/home (accessed June 2016)

2 Related Work

In the previous chapter, we introduced *geographic information retrieval* (GIR) as a research field which provides automatic methods for the retrieval of spatial, temporal, and thematic information from unstructured and semi-structured contents. Therefore, relevant work using GIR in the context of this thesis is reviewed in *Section 2.1*. The *spatialization framework*, applied for the retrieved spatial, temporal, and thematic data in this project is described in *Section 2.2*. For the incorporation of spatialized displays in dynamic and interactive web interfaces, as well as for the design and the evaluation of these interfaces, *geovisual analytics* was found to be relevant and is therefore illustrated in *Section 2.3*. The results of this project should serve the *digital humanities* community. Therefore, this research field is presented in *Section 2.4*. The research gap identified by analyzing previous work related to this thesis in *Sections 2.1-2.4* is outlined in *Section 2.5*.

2.1 Geographic information retrieval

A common assertion in GIScience is that 80% of all information has a reference to space or geography (e.g., MacEachren and Kraak, 2001: 3). Although it is difficult to identify the origin of this assertion in the literature, Hahmann et al. (2011) traces it back to the 1980s and discovered that initiators primarily used it to describe municipal and governmental data and thus referred to coordinates, identifiers, and address data. In Hahmann and Burghardt (2013), the authors highlight the difficulty in generally defining what exactly *an information with a geographic reference* is, and test this assertion with their own definition of *geospatially referenced information* in another data source context (i.e., user-generated data from *Wikipedia*). As an outcome of this study, Hahmann and Burghardt (2013: 1171) question the general validity of the assertion and conclude that for the corpus they investigated (i.e., *Wikipedia*) they could not confirm the *80%-assertion*, but rather argue for a *60%-assertion*. Adams and Gahegan (2016) also studied spatial information on *Wikipedia*, but used a slightly different approach than Hahmann and Burghardt (2013). They found that 54.3% of all articles on *Wikipedia* contain at least one reference to a populated place, and 74.7% of all articles on *Wikipedia* refer to any place type (e.g., populated places, museums, parks). Although the results of the research projects by Hahmann and Burghardt (2013) and by Adams and Gahegan (2016) do not confirm the *80%-assertion*, the current prevalence of space and geography in information is indisputable, and massive amounts of geographic information are stored online; for

example, in computer databases, digital maps and web pages, including text (e.g., articles, books, reports), and images containing geographic information (Jones and Purves, 2008: 219). The automatic retrieval of geographic information from unstructured or partly structured (online) data has challenged researchers in forming a new scientific field coined *geographic information retrieval* (GIR).

Research in GIR extends existing approaches in *information retrieval* (IR). Sanderson and Croft (2012: 1444) stated that "an IR system typically searches in collections of unstructured or semi-structured data (e.g., web pages, documents, images, video, etc.)", and "(…) locates information that is relevant to a user's query", and is needed "(…) when a collection reaches a size where traditional cataloguing techniques can no longer cope". Web search engines (e.g., *Google*[15], *Bing*[16], *Yahoo!*[17]) are typical examples of IR systems. Users enter a query and the most relevant web search results are usually listed and presented in a ranked order to the user. Although research in IR dates back further than the origin of web search engines (the early 1990s), the current growth of data available online has spurred an interest and need for new approaches in IR (Sanderson and Croft, 2012: 1448). As a result, IR systems have become highly complex and involve, for example, (web) link analysis and analysis of anchor text in hyperlinks (e.g., Brin and Page, 1998), user query log data (e.g., Radlinski and Joachims, 2005), as well as probabilistic approaches and language models (e.g., Hiemstra, 1998, Ponte and Croft, 1998). Based on these IR developments in recent years, GIR has begun to address the research challenges which are particularly relevant to geographic information (Jones and Purves, 2008: 219). For example, place names must be disambiguated (e.g., does *London* refer to *London, UK* or *London, Ontario, Canada*?) in order to determine which particular instance of a name is implied in a data source (Jones and Purves, 2008: 220). Furthermore, having retrieved geographic information, it needs to be incorporated appropriately into user interfaces with geographic search options (Jones and Purves, 2008: 222-23). These and further challenges have been addressed by the GIR community and are covered in the following subsections. In this thesis, we focus on GIR methods for digital text data in retrieving geographic information from unstructured and semi-structured text documents.

Although much emphasis in GIR has been placed on the spatial dimension of geographic information, temporal and thematic information is important in the search and retrieval of (geographic) information as well, as recently highlighted by Ballatore et al. (2015: 12) and Grossner (2014). Similarly, the three dimensions of geographic information (i.e., spatial, temporal, and thematic) are important in the context of our project, as we aim at incorporating all of them in spatialized displays and interactive web interfaces with spatio-temporal and thematic information search functionalities, as described in *Section 1.3*. Therefore, in *Subsections 2.1.1-2.1.3*, the retrieval of spatial, temporal, and thematic information from unstructured and semi-structured text documents is covered. In *Subsection 2.1.4*, GIR systems and evaluation methods are

[15] Google web search engine: https://www.google.com (accessed April 2016)
[16] Bing web search engine: https://www.bing.com (accessed April 2016)
[17] Yahoo! web search engine: https://www.yahoo.com (accessed April 2016)

briefly illustrated. In the following subsection, we first focus on the retrieval of spatial information.

2.1.1 Retrieving spatial information

Only in recent years has much effort been placed toward the development of computer systems for retrieving spatial information from unstructured and semi-structured digital text data (Jones and Purves, 2009). An early and influential work in this field was published by Larson (1996), which reported on techniques and problems related to GIR and *spatial browsing* in *digital libraries*. Later, Sanderson and Kohler (2004), Gan et al. (2008), and Jones et al. (2008) studied the importance of spatial information in web queries and reported that approximately 12-15% of queries sent to traditional web search engines contain a place name. This further highlights the need for research on how to retrieve spatial information from online content and how to incorporate it into modern IR systems. The GIR community has elaborated on answers to this challenge, and relevant research in the context of this thesis is summarized in this subsection. A discussion of relevant research is structured following the GIR model for digital text data presented by Leidner and Lieberman (2011), and is illustrated in Figure 3.

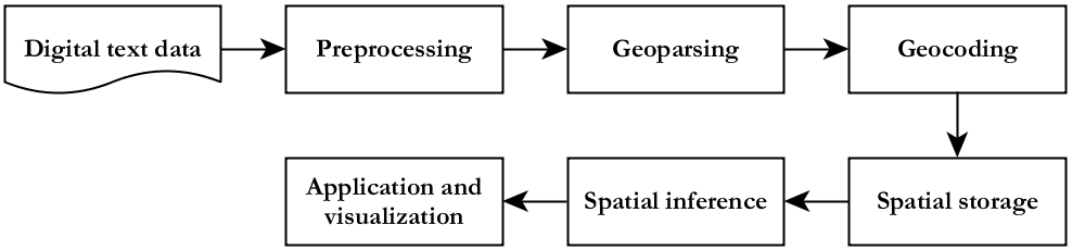


Figure 3: Model for retrieving spatial information from unstructured or semi-structured digital text data (modified from Leidner and Lieberman, 2011: 6).

Preprocessing

During preprocessing raw text is extracted from the original data source, which could be a digital text archive or a collection of web pages containing text data, images, and videos embedded in HTML[18] files (Leidner and Lieberman, 2011: 6). An example of this could be an online archive of a newspaper, from which several thousands of news articles in HTML format are retrieved. From these articles, all non-textual elements (e.g., images, videos, HTML tags) are removed in order to work with raw text in further stages of the GIR process.

Geoparsing

In this step, all occurrences of place names (i.e., toponyms) or descriptions of places and their respective type of geographic feature (e.g., cities, villages, rivers, mountains) in each

[18] HyperText Markup Language (HTML): https://www.w3.org/html/ (accessed April 2016)

of the text documents are detected (Leidner and Lieberman, 2011: 6). In the context of a newspaper, this would imply that in an article about the city of *London, UK*, all occurrences of *London* and all other toponyms mentioned in the article (e.g., the river *Thames*, the city *Watford* northwest of London) have to be identified, retrieved, and classified (i.e., *London* = city) by a GIR system. Various approaches used to complete this step are categorized by Leidner and Lieberman (2011: 7-8) as follows.

- **Gazetteer lookup based**. Texts are processed word by word or character by character and compared to a list (i.e., a gazetteer) with toponyms and associated metadata (e.g., Hill et al., 1999, Hill, 2006). In the previously mentioned newspaper example, this implies we might employ a gazetteer consisting of all toponyms (e.g., cities, rivers, mountains) in the UK, given that we are interested in retrieving toponyms from the UK only. Then, a gazetteer based GIR system processes all texts and identifies *potential toponyms*. Therefore, in our example using the *London, UK* newspaper article, we would expect the GIR system to detect toponyms such as *London*, *Thames*, and *Watford* as *potential toponyms* as they all occur in the selected UK gazetteer. As illustrated by Mikheev et al. (1999), gazetteer based approaches often work effectively. An example of a typically employed gazetteer is *GeoNames*[19], which contains more than eight million toponyms from all over the world, and is available for download free of charge. However, if working with gazetteers, the coverage in a study area must be critically assessed, as the global coverage of gazetteers such as *GeoNames* is not a simple mirror of population density and is not evenly distributed among regions and countries (Graham and De Sabbata, 2015). *GeoNames* is constructed from national gazetteers and data sets as well as user-generated content[20] which results in diverging national and international policies as well as a geographically uneven distribution of contributors to user-generated content, ultimately resulting in unequal global coverage (Graham and De Sabbata, 2015).

- **Rule based**. Rules are defined for a GIR system to locate specific structures (e.g., grammatical structures) in text documents in order to identify *potential toponyms* (Leidner and Lieberman, 2011: 7-8). In the previous newspaper example, an algorithm could, for example, be employed which searches for all occurrences of *city of* in the text documents, and if *city of* is detected, the subsequent term(s) is/are a *potential toponym*. Therefore, in the following sentence, *London* and *Watford* should be recognized as *potential toponyms*: "The *city of London* and the *city of Watford* are located in the UK." The use of *regular expressions*[21] is one possibility to implement such rules. Whereas Thompson (1968) provided initial work on *regular expressions*, Friedl (2006) investigates how to master *regular expressions* in modern text and data manipulation tasks. To summarize these works, *regular expressions* consist of a sequence of characters defining a search pattern. For example, the search pattern *city\sof\s\w+* finds all occurrences of *city of*, followed by a word as the expression *\s* matches

[19] GeoNames: http://www.geonames.org/ (accessed April 2016)
[20] GeoNames data sources: http://www.geonames.org/data-sources.html (accessed April 2016)
[21] Detailed information about Regular Expressions: http://www.regular-expressions.info/ (accessed April 2016)

whitespace characters, whereas \w+ matches a sequence of word characters. Therefore, expressions such as *city of London* or *city of Watford* match this *regular expression* pattern. Mikheev et al. (1999) report that only considering rule based approaches for GIR does not produce satisfying results compared to approaches which also incorporate gazetteer lookup.

- **Machine learning based**. At each position in the text, a pre-defined set of properties (i.e., *features*) is computed (Leidner and Lieberman, 2011: 8). *Features* usually incorporate checks for particular strings, length computations, and capitalizations, and are implemented as *Boolean tests* (i.e., true or false) in a GIR system (Leidner and Lieberman, 2011: 8). In a *test corpus* (i.e., raw text), all *features* in the text documents are computed and then compared to feature configurations which were retrieved from a *training corpus* containing *gold data*[22]. Based on this comparison (e.g., through *statistical inference*), the GIR system decides if a term or several terms is/are a *potential toponym* (Leidner and Lieberman, 2011: 8). Translated to the prior newspaper example, this implies that a GIR system may identify all occurrences of the term *river* including one subsequent word. Therefore, the expression *river Thames* in a text would be assigned the value *true* for this feature (i.e., *river* and *subsequent word*). Then, this feature is statistically correlated to the target outcome during training on the *gold data* corpus (Leidner and Lieberman, 2011: 8). This could lead to a high probability that *Thames* is a *potential toponym* as this feature (i.e., the word *river* followed by a toponym) is very common in the *training corpus*. As a result, *Thames* would be recognized as a *potential toponym*.

Typically, several *geoparsing* categories are combined in a GIR system to optimize the retrieval results. For example, Li et al. (2002) and Li et al. (2003) propagated a gazetteer based approach which was followed by a rule based approach, whereas Martins et al. (2010) and Santos et al. (2014) used a combination of gazetteer lookup and machine learning approaches.

Geocoding

We have described that *potential toponyms* are detected in the *geoparsing* step. In the *geocoding* step, ambiguity issues among the *potential toponyms* are resolved and geographic coordinates are assigned to the resolved toponyms. Amitay et al. (2004) categorized the ambiguity of toponyms as follows.

- **Geo/non-geo ambiguity**. This ambiguity arises if a place name also has a non-geographic meaning (e.g., *Turkey* equals a country name as well as an animal, and *London* equals a city name as well as a last name). One special case of this category is metonyms (i.e., *Washington* is either used as a toponym or in reference to the US government). The importance of metonyms in *geocoding* was demonstrated by Leveling and Veiel (2007: 902), who studied a German newspaper corpus and reported that 17% of location names are used metonymically.

[22] Gold data: A reference corpus, in which all occurrences of geographic names or phrases have been manually annotated (Leidner and Lieberman, 2011: 8).

- **Geo/geo ambiguity**. This ambiguity refers to different places with the same name (e.g., *London, UK* and *London, Ontario, Canada*). The number of toponyms with *geo/geo ambiguity* in texts heavily depends on the studied data set. Amitay et al. (2004: 274) reported 37% *geo/geo ambiguity* on web pages, whereas Smith and Crane (2001) identified 92% of the toponyms in a digital history library as potentially referring to more than one location.

Often *geo/non-geo ambiguities* are already resolved in the *geoparsing* stage by applying rule based gazetteer approaches. If not, they could be resolved in the *geocoding* stage by measuring the frequency of detected *potential toponyms* in large reference corpora. This helps to assess how common the *potential toponyms* are in standard language. Amitay et al. (2004) used 1,200,000 pages from the .gov domain as reference corpus whereas Derungs and Purves (2014) employed web counts using the *Yahoo! API*. The higher the frequency of *potential toponyms* in the reference corpora, the higher the authors weighted the probability of *geo/non-geo ambiguity* of *potential toponyms*. As a result, *potential toponyms* with a high probability of *geo/non-geo ambiguity* are excluded in order to reduce the number of incorrectly identified toponyms.

A common used approach to resolve *geo/geo ambiguity* employs the definition of a *default location* associated with location metadata (Purves et al., 2007: 726). Possible metadata to determine *default locations* are population counts (e.g., Rauch et al., 2003, Amitay et al., 2004) or the importance of places (e.g., Pouliquen et al., 2006). In the first case, to disambiguate *London* in a text, the population counts of places called *London* (*London, UK*; *London, Ontario, Canada*, etc.) are considered, and the *potential toponym* is referred to the place which has the highest population (i.e., *London* to *London, UK*). In the latter case, the importance of places called *London* is considered (e.g., higher importance of *capital city* compared to a *non-capital city*) and the *potential toponym* is referred to the place which has the highest importance. According to this approach, *London* is rather referred to *London, UK* than to *London, Ontario, Canada*, as *London, UK* is a capital, whereas *London, Ontario, Canada* is not.

Another common approach to resolve *geo/geo ambiguity* incorporates textual and geographic distance. Rauch et al. (2003: 52) motivate the use of this approach by highlighting that "(…) there is a high degree of spatial correlation in geographic references that are in textual proximity". Therefore, the closer an unambiguous toponym occurs to an ambiguous *potential toponym* in the text, the more relevant this unambiguous toponym is considered in disambiguating the *potential toponym*. We illustrate this with the previous *London* example: if unambiguous toponyms occur in close textual proximity (e.g., in the same sentence) to *London* in a text, and these unambiguous toponyms are all located in the UK, the *potential toponym* is rather resolved to *London, UK* than to *London, Ontario, Canada*. This procedure reflects Tobler's *first law of geography* (Tobler, 1970: 236) applied to texts, claiming that "everything is related to everything else, but near things are more related than distant things". The popularity of this approach is evident in its application in the work of Rauch et al. (2003), Pouliquen et al. (2006), and Derungs and Purves (2014).

In the existing literature, a large number of disambiguation methods are described: for example, interested readers are referred to Leidner (2007) and Buscaldi (2011), which provide an overview of existing methods. Multiple approaches, as opposed to only one disambiguation method, are typically considered and combined in the *geocoding* process (e.g., Derungs and Purves, 2014).

Spatial storage and spatial inference

The next stages after *geocoding* are *spatial storage* and *spatial inference*, according to the model presented in Figure 3. *Spatial storage* refers to the process of storing the extracted spatial information in an efficient data structure in order to enable fast retrieval of the information (Leidner and Lieberman, 2011: 7). In traditional IR systems, words are indexed, and inverted indexes[23] are used to efficiently locate matching documents (Zobel and Moffat, 2006). As an alternative or supplement to these approaches, in GIR, more sophisticated methods in which index elements correspond to *toponyms*, *toponym identifiers*, *spatial footprints* (i.e., coordinates), or complex shapes such as *minimum bounding rectangles* and *polygons* are propagated (Leveling, 2011).

Spatial inference refers to performing *reasoning operations* based on *spatial logic* (Leidner and Lieberman, 2011: 7). *Transitivity rules* might also be applied for *reasoning operations* which implies that if, for example, toponym *A* is south of *B*, and *B* is south of *C*, then *A* is as well south of *C*.

Application and visualization

The final stage in Figure 3 refers to the implementation of applications which enable access to extracted spatial knowledge (Leidner and Lieberman, 2011: 7). Normally, GIR systems present retrieved results as locations or clusters on a map (e.g., Leidner et al., 2003).

In this subsection, we have illustrated various approaches used to retrieve spatial information from unstructured and semi-structured text documents based on Leidner and Lieberman's (2011) spatial information retrieval model, as illustrated in Figure 3. In the next subsection, we focus on the retrieval of the second dimension of geographic information: time.

2.1.2 Retrieving temporal information

The value and the importance of temporal information has been recognized by the IR community, particularly in recent years. Alonso et al. (2007, 2011) pointed out the huge and only scarcely exploited potential of temporal information for providing alternative search features (e.g., document exploration, similarity search, summarization, clustering), and enhancing user experience in IR systems. However, the retrieval of temporal information remains a major research challenge in IR, as information regarding time is

[23] Inverted index: mapping of unique words to documents or a set of documents in which the unique words appear. See Zobel and Moffat (2006) for further information.

often *implicit* (e.g., *Christmas 2016*), *vague* (e.g., *several days later*), or *underspecified* (e.g., *December*) in unstructured or semi-structured web content (Derczynski et al., 2015: 786).

The IR community dealing with the retrieval of temporal data has proposed various approaches aimed to retrieve temporal information from unstructured or semi-structured web content. In the following paragraphs, approaches relevant to the context of this thesis are summarized. The discussion is structured following the temporal IR model suggested by Strötgen and Gertz (2013), which is illustrated in Figure 4. The preprocessing step is not discussed, as it is similar to the preprocessing step explained in *Subsection 2.1.1* for spatial information retrieval.

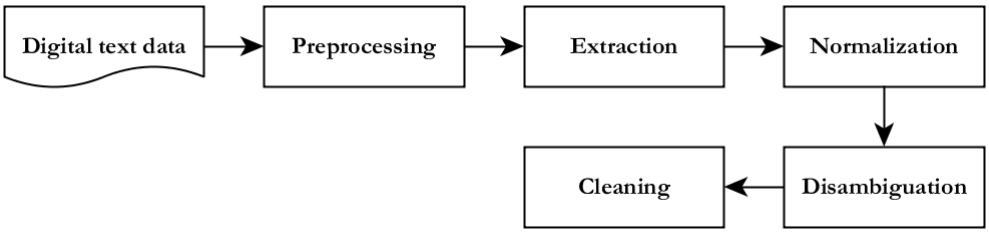


Figure 4: Model for retrieving temporal information from unstructured or semi-structured digital text data (Strötgen and Gertz, 2013: 287).

Extraction

Schilder and Habel (2001) and Alonso et al. (2007) distinguished between three different categories of temporal information: *explicit* (e.g., *September 27, 2016*), *implicit* (e.g., *Christmas 2016*), and *relative* (e.g., *today*) expressions of time. Whereas *explicit* expressions can be extracted using existing *part-of-speech taggers*[24] (e.g., Schmid, 1994), the retrieval of *implicit* and *relative* expressions of time requires more sophisticated IR methods. For these categories, advanced *rule based* (e.g., Strötgen and Gertz, 2013) and *machine learning* strategies are suggested in the literature. A common element of rule based approaches is *regular expressions* (see *Subsection 2.1.1*). For example, *1990-95* matches the *regular expression* *\d\d\d\d-\d\d* in a text as *\d* is the *regular expression* for a single digit. *Machine learning* approaches are not further discussed here as they are irrelevant to the context of this thesis. However, interested readers are referred to Hacioglu et al. (2005) and Ahn et al. (2007) for further clarification on *machine learning* techniques.

Normalization

In the *normalization* stage, the extracted expressions of time are converted to a standard format, meaning that expressions referring to the same point in time are assigned the same value (Strötgen and Gertz, 2013: 274). Currently, two major temporal standards are employed in temporal IR: *Tides Timex2* and *TimeML*. *TimeML* is an extension of the *Tides Timex2* standard, and is covered in the following paragraphs. However, readers interested in *Tides Timex2* are referred to Ferro et al. (2001, 2005) for further clarification.

[24] Part-of-speech taggers annotate category of words (e.g., nouns, verbs, numerals) in texts.

TimeML consists of guidelines for the annotation of temporal expressions according to the *ISO 8601 standard*[25], with some extensions. *TimeML* uses *Timex3* tags to annotate temporal references in texts (Pustejovsky et al., 2005). *Timex3* distinguishes between four tag types: *date* (e.g., *September 27, 2016*), *time* (e.g., *5 p.m.*), *duration* (e.g., he worked for the company *for five years*), and *set* (e.g., *weekly*) (Pustejovsky et al., 2005: 145). For example, the date *September 27, 2016* is normalized to *2016-09-27*, the time *5 p.m.* to *XXXX-XX-XXT17:00*, the duration *for five years* to *P5Y*, and the set *weekly* to *P1W*. Almost all temporal taggers normalize temporal expressions employing *rule based* approaches (Strötgen and Gertz, 2013: 278).

Disambiguation

In the *disambiguation* stage, *ambiguous* and *underspecified* temporal expressions are resolved (Strötgen and Gertz, 2013: 287-88). We illustrate an example for resolving ambiguity with the temporal expression *September, 27, 2016*: in the *extraction* and the *normalization* stage, (1) *September 27, 2016*, (2) *September, 27*, (3) *September*, and (4) *2016* are identified as temporal references for the expression *September, 27, 2016* (Strötgen and Gertz, 2013: 287). Therefore, in the *disambiguation* stage, all but the longest expression (i.e., (1) *September 27, 2016*) have to be removed.

The expression *5 p.m.* which is normalized to *XXXX-XX-XXT17:00* is an example of an *underspecified* temporal expression since *year*, *month*, and *date* are missing. In order to specify this information, the temporal expression previous or subsequent to *5 p.m.* in the text (if available) is assumed to be the reference time for narrative texts, as demonstrated in Strötgen and Gertz (2013: 288). To illustrate this, we use the following example: "We went for a hike on *September 27, 2016*. After the hike, we took the train at *5 p.m.* to return back home". *September 27, 2016* is assumed to be the reference date for *5 p.m.*, as it occurs in the first sentence and thus the normalized value for the temporal reference *5 p.m.* in the second sentence is transformed from *XXXX-XX-XXT17:00* to *2016-09-27T17:00*.

Cleaning

In the cleaning phase, all expressions which were identified as potential temporal expressions in the *extraction* stage, but identified as *non-temporal* and therefore invalid in the *normalization* stage, are removed (Strötgen and Gertz, 2013: 288). For example, in the *extraction* stage, the expression *1990 miles* is identified as potentially referring to the year *1990*. However, in the *normalization* stage of common *rule based* temporal IR systems, numbers which are followed by a plural noun are identified as *non-temporal expressions* and are thereby marked for removal in the *cleaning* stage (e.g., in Strötgen and Gertz, 2013: 287). Strötgen and Gertz (2013: 285-86) highlight the risk of falsely removing expressions which are temporal (e.g., *the 2000 celebrations*) by applying this rule, and state that resolving this and similar issues is a current research challenge in temporal IR.

[25] ISO 8601 date and time format: http://www.iso.org/iso/home/standards/iso8601.htm (accessed October 2016)

In this subsection, we have illustrated the retrieval of temporal information from unstructured and semi-structured text documents based on the temporal information retrieval model by Strötgen and Gertz (2013), which is illustrated in Figure 4. In the following subsection, we focus on the retrieval of the third dimension of geographic information: theme.

2.1.3 Retrieving thematic information

Retrieving thematic information from unstructured and semi-structured text documents is another field of research in IR. Steyvers and Griffiths (2008) distinguish between *content* and *form based* approaches, which are illustrated in Figure 5. *Content based* approaches seek to generalize and conceptualize the entire content of single text documents (e.g., assigning topics to text documents). In contrast, *form based* approaches focus on single terms (e.g., keywords) or a combination of terms in text documents. A very simple example of a *form based* approach is to calculate the *frequency of words* in text documents. If a user is interested in a specific term, the user could enter the term in a *form based* IR search engine and the text document in which the search term occurs most often is presented to the user as the most relevant search result.

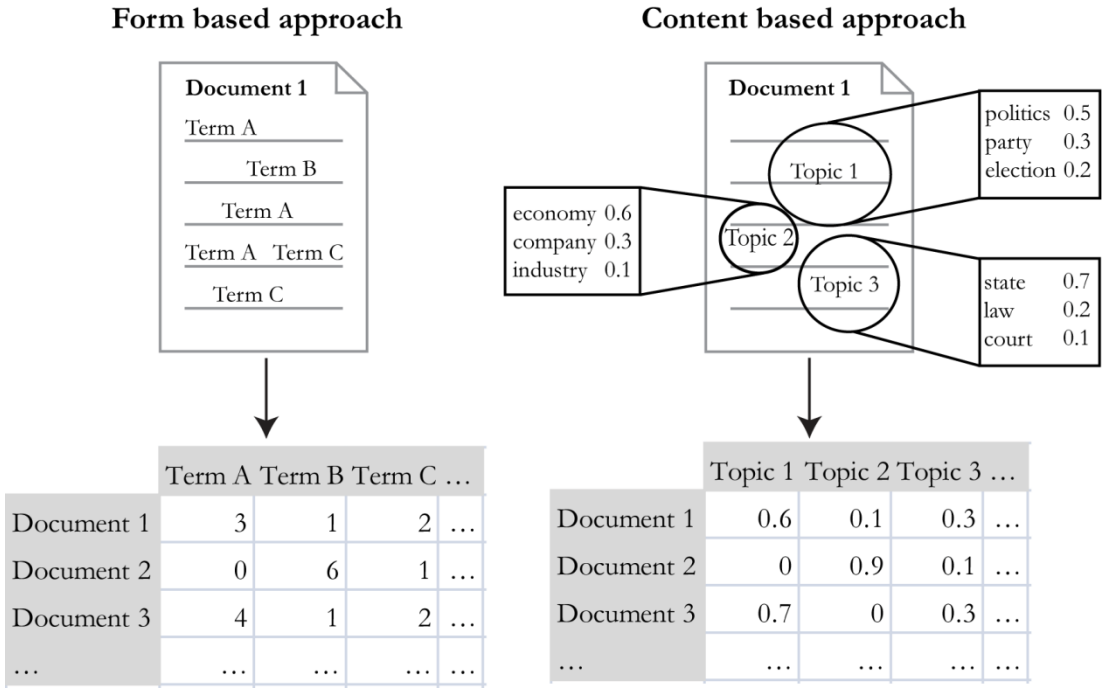


Figure 5: Comparison of *form based* (left) and *content based* (right) approaches to retrieve thematic information in IR.

The *form based* approach is exemplified in Figure 5 (left). In each row of the table, the frequencies of *Terms A*, *B*, and *C* in *Documents 1*, *2*, and *3* are shown. If a user enters *Term A* in a *frequency* only based IR search engine, *Document 3* and *Document 1* are most relevant and therefore returned to the user as they both contain *Term A* several times. *Document 3* is considered more relevant than *Document 1* because *Term A* occurs once more in *Document 3* compared to *Document 1*. Many extensions of the frequency only approach exist; for example, the *term frequency-inverse document frequency* (tf-idf) weighting

scheme (Manning et al., 2009d), which is very common in IR. The *tf-idf* considers not only the frequency of terms in documents to describe the documents, but also includes characteristics of the entire corpus. As a result, terms which are very rare in the entire corpus are assumed to be more relevant to describe a single document than very common words in the corpus, assuming that both terms have the same frequency. In other words, if both the terms *salt* and *coriander* occur in a recipe of an Italian cookbook, *coriander* is assumed to be more specific for the recipe than *salt*, as it occurs less often in a typical Italian cookbook than *salt*. A further extension is the *Okapi BM25* (e.g., Robertson and Zaragoza, 2009), which is discussed in detail in *Subsection 4.2.1* of this thesis.

The *frequency* or *tf-idf* values of terms in a document are usually represented in a *vector space model* (VSM) (Dubin, 2004). In the VSM, each document of a corpus is a vector and each dimension of the vector represents the *frequency* or the *tf-idf* value of a term. The similarity of documents is calculated, for example, by applying the *cosine similarity* measure, which calculates the cosine of the angle between vectors. In a similar vein, the similarity of a query and a document can be calculated. This implies that if a user enters a query in a *form based* search engine (e.g., *frequency* or *tf-idf*), the vector of the query and the vector of the documents are compared. Then, the document-query combination which has the highest *cosine similarity* value is assumed to be most relevant and thus presented to the user in a search engine.

In contrast, *content based* approaches generalize the content of documents on a conceptual level (Steyvers and Griffiths, 2008) instead of assigning values to individual words. *Probabilistic topic modeling* is one very commonly used *content based* approach to assess semantic themes in documents (Steyvers and Griffiths, 2007). *Probabilistic topic modeling* originated from *latent semantic indexing* (LSI), which is also referred to as *latent semantic analysis* (LSA) in the literature (e.g., Deerwester et al., 1990, Landauer and Dumais, 1997, Dumais, 2004). LSI is a typical dimension reduction technique. Words which occur often together in documents are summarized in *concepts*, so that documents are represented in fewer dimensions than unique words exist in the documents.

The first two stages of a typical LSI procedure are identical to *form based* approaches: a *term-document matrix* is calculated and raw term frequency values are weighted, for example, using the *tf-idf* method (Dumais, 2004: 192). As an extension to the *form based* approach, dimensions are subsequently reduced by applying the *singular value decomposition* (SVD) technique (Golub and Reinsch, 1970), which is similar to *principal components analysis* (e.g., Jolliffe, 2002). In short, SVD summarizes words which often occur together in *concepts* and optimizes the result in order to preserve the similarity structures among documents in the corpus (Dumais, 2004). For example, the terms *politics*, *party*, and *election* in Figure 5 (right) might be combined to the general concept *politics*, as these words often occur together in texts about *politics*. Interested readers are referred to Golub and Reinsch (1970) and Stewart (1993) for further details about the SVD.

The principle of LSI was further elaborated by Hofmann (1999), who presented the *probabilistic latent semantic indexing* (pLSI) and by Blei et al. (2003), who introduced the

latent dirichlet allocation (LDA) technique. Both pLSI and LDA extend standard LSI techniques by incorporating probability distributions. Documents are modeled as a probabilistic mixture of topics, and a topic consists of a probability distribution over words. This is illustrated in Figure 5 (right). *Document 1* consists of three topics. *Topic 1* contains the terms *politics*, *party*, and *election* with probability values of 0.5, 0.3, and 0.2, respectively. This implies that the term *politics* is most descriptive, *party* is second most descriptive and *election* is least descriptive of *Topic 1*.

In the table within Figure 5 (right), *Topic 1* is assigned a value of 0.6 for *Document 1*, which implies that *Topic 1* is most descriptive for *Document 1* compared to *Topic 2* and *Topic 3* with values of 0.1 and 0.3, respectively. Equal to the *form based* approach, the similarity of documents can be calculated by applying *cosine similarity* measures on the document-topic vectors. *Document 1* and *Document 3* in Figure 5 (right) consist of a similar topic mixture distribution, and, consequently, the *cosine similarity* of *Document 1* and *Document 3* is higher (i.e., the documents are more similar) than the *cosine similarity* of *Document 1* and *Document 2*, or *Document 2* and *Document 3*.

Both *form based* and *content based* approaches are applied in IR systems. Sometimes, the approaches are combined in order to optimize retrieval results (e.g., Steyvers and Griffiths, 2008).

2.1.4 GIR systems and evaluation

In the previous subsections, we have illustrated the retrieval of all three dimensions of geographic information (i.e., space, time, and theme). In this subsection, we focus on GIR systems and evaluation methods. In GIR systems, all stages, from raw text data processing to the retrieval and presentation of query results to an interested information seeker, are integrated in a holistic solution. GIR systems support the retrieval of geographic information from unstructured or semi-structured digital text corpora in different ways. In this subsection, one example of a well-known GIR system and an extension of this system are presented, and methods to evaluate GIR approaches are briefly discussed.

GIR systems

Spatially-Aware Information Retrieval on the Internet (SPIRIT) is a GIR system which integrates *spatial* and *thematic*[26] information in a holistic way (Purves et al., 2007). In addition, spatial relationships (e.g., *near*, *inside*, *north of* a specified location) in text documents are considered (Purves et al., 2007: 731). Therefore, the system resolves queries in the form of <theme><spatial relationship><location> (Purves et al., 2007: 723). In the example query, *castles in Wales* the *theme* is *castles*, the *spatial relationship* is *in*, and the *location* is *Wales*. Furthermore, some imprecise regions (e.g., *Swiss Mittelland*, *American Midwest*) are integrated as locations in the SPIRIT search engine (Purves et al., 2007: 728).

[26] In this context, *thematic* describes the non-spatial content of a query (e.g., *castles* in the query *castles near Southampton*).

In the *geoparsing* stage, a combination of *gazetteer lookup* and *rule based* approaches were applied in order to identify *potential place names* (Purves et al., 2007: 725). Furthermore, proper names and commonly occurring terms which are most likely used in a non-geographical sense were excluded (Purves et al., 2007: 726). In the *geocoding* stage, potentially ambiguous place names were resolved using a *default location* approach (see *Subsection 2.1.1*) (Purves et al., 2007: 726). In the next stage, disambiguated place names were stored in a hybrid (i.e., spatial and thematic) index structure (Purves et al., 2007: 727).

Figure 6: SPIRIT search engine (Bucher et al., 2005).

A sample search in SPIRIT is depicted in Figure 6. The spatial dimension of a search query can either be defined by entering place names in a text field or by manually drawing a region directly in the map. The ranked retrieval results are presented on the map (left) and as a document list (right). The ranking is based on thematic as well as spatial relevance (Purves et al., 2007: 730). In short, relevant documents need to match with the *theme* (i.e., *castles* in Figure 6), the *location* (i.e., *Wales* in Figure 6), and the *spatial relationship* (i.e., *In* in Figure 6) specified in the query. Further details regarding the SPIRIT system are described in Purves et al. (2007).

The PIV system (Palacio et al., 2011) conceptually extends SPIRIT as it involves all geographic dimensions (i.e., *space*, *time*, and *theme*). Therefore, relevant results for a query such as *Potato famine in Ireland after the mid-19th century* are retrieved based on *spatial* (e.g., *Ireland*), *temporal* (e.g., *mid-19th century*), and *thematic* (e.g., *potato famine*) information in text documents (Palacio et al., 2011: 92). More details regarding the PIV system are provided in Palacio et al. (2011).

Evaluation

GIR evaluation is often divided into *system-oriented* and *user-oriented* approaches (Ingwersen and Järvelin, 2005). *System-oriented* approaches evaluate the IR algorithms and its outputs, and thus how well relevant documents are retrieved and distinguished from irrelevant documents, and how well the ranking algorithms perform without involving target users in the evaluation process (Bucher et al., 2005). In contrast, *user-oriented* approaches evaluate the IR system or specific components of the IR system in order to assess how well it supports the information-seeking processes, involving target users (Borlund, 2009, Kelly, 2009).

System-oriented approaches are based on a set of performance measures which have been described by Manning et al. (2009b), of which *recall* and *precision* are the best known measures. *Precision* expresses the fraction of relevant documents out of all retrieved documents by an IR system, whereas *recall* expresses the fraction of relevant documents which are retrieved out of all relevant documents. For example, if 100 documents are retrieved from a corpus and 70 of them are relevant, the *precision* is 0.7. In contrast, if 80 relevant documents exist in a corpus and 50 of these relevant documents are retrieved by an IR system the *recall* is 0.625, regardless of the number of retrieved irrelevant documents. Both measures disregard the ranking of documents in an IR system output. Measures which regard the ranking of documents are, for example, *(Mean) Average Precision* and *Normalized Discounted Cumulative Gain*, which are not discussed in detail in this thesis; however, interested readers are referred to Manning et al. (2009b).

To compare different IR systems and retrieval algorithms following a *system-oriented* approach, *test collections* are typically employed (Sanderson, 2010). These *test collections* consist of domain-specific corpora and queries representing users' information needs (Peters and Braschler, 2001). The relevance of (retrieved) documents in the corpora is manually evaluated for the specified queries by people who read the documents and decide upon their relevance to a specific query (Peters and Braschler, 2001). IR algorithms are run on these *test collections*, and performance measures of various IR systems or configurations of a single IR system are compared to one another (Sanderson, 2010: 251). *Test collections* are discussed in detail in Sanderson (2010).

If a *user-oriented* approach is applied, target users of an IR system test the system in user studies by solving given tasks (Kelly, 2009). In these user studies, common *usability* values (e.g., *task success*, *time on task*, *numbers of errors*, *number of documents viewed*) are assessed (Kelly, 2009). Furthermore, users provide answers to *System Usability Scales* (Brooke, 1996) regarding the *ease of use* of an IR system, and about their opinion if they were

successful in completing the tasks (Kelly, 2009). In addition, users are asked to judge the relevance of retrieval results (Park, 1994).

In GIR, *system-oriented* evaluation strategies are widespread (Bucher et al., 2005). For example, the evaluation of the PIV engine is *system-oriented* (Palacio et al., 2011). As a focus on target users in GIR has been highlighted recently (e.g., Mandl, 2011), combined approaches (i.e., *system-oriented* and *user-oriented*) have been developed and applied as, for example, in the evaluation of the SPIRIT system (Bucher et al., 2005).

Summary

- GIR systems support geographic search tasks in large digital text archives, as they identify geographic information (i.e., spatial, temporal, and thematic information) in unstructured and semi-structured text documents and provide functionalities to search for and access spatial, temporal, and thematic information in text documents (e.g., search engines).
- GIR provides methods to disambiguate geographic information in text documents (e.g., does *London* refer to *London, UK* or *London, Ontario, Canada*?) and further suggests evaluation methods to test how well information retrieval algorithms perform.

For this thesis, the GIR approaches illustrated in this section are relevant to answering *Research Question 1*, as we aim to automatically retrieve geographic information from a semi-structured text archive (see *Section 1.3*). Furthermore, evaluation methods suggested by the GIR community are relevant in order to test the results obtained in this thesis, as illustrated in *Chapter 6 – Evaluation*.

The next step of our research approach concerns the depiction of retrieved geographic information in visualizations that highlight spatio-temporal and thematic structures as well as interconnections in text data (see *Section 1.3*). This complies with Moretti's (2005) *distant reading* approach to reduce and abstract text documents in order to assess their overall interconnections, shapes, relations, and structures. Moretti (2005) suggested the use of *graphs*, *maps*, and *trees* to depict reduced and abstracted information in texts. We consider the *spatialization framework* for this research project, which proposes similar visual means as demonstrated in the following section.

2.2 Spatialization

The *spatialization framework* is a theoretical framework including practical implications to generalize, abstract, and depict large multidimensional data archives in spatialized displays. Spatialized displays are based on *spatial metaphors* and aim at supporting information seekers discovering structures and interconnections in the data, and to gain new insights, which is detailed in the current section.

Due to the rapid technological advances in computer and communication technologies since the 1970s, the amount and size of digital data archives have increased exponentially. As a result, the interest of an interdisciplinary research community has been attracted to find solutions which facilitate access to information in large data archives (Fabrikant, 2000). *Computer scientists*, *data mining* experts, *knowledge discovery* specialists, and researchers in related fields have developed computational techniques and employed statistical analyses to address this challenge (Fayyad et al., 1996). Experts in *information visualization*, *psychology*, and *cognitive science* have been interested in computational approaches as well, but have paid particular attention to the cognitively adequate presentation of information extracted from large data archives by visual means (e.g., Fabrikant, 2000, Keim, 2002).

Information visualization suggests solutions to reduce rich and multidimensional information to lower dimensional visualizations that facilitate sense-making by information seekers, since humans possess a limited information-processing capacity (Keim, 2002, Fabrikant and Skupin, 2005). The transformation from multidimensional to lower dimensional representations is based on the use of *(spatial) metaphors* (Fabrikant and Buttenfield, 2001). Metaphors support humans in understanding complex theories and models. The expression *life is a journey* is an example of this. The term *journey* is used as a metaphorical expression to describe the meaning of *life* (Lakoff, 1987). Metaphors are mappings from a *source* (i.e., *journey*) to a *target domain* (i.e., *life*) in order to better understand the *source domain* (Lakoff and Johnson, 1980). This idea of metaphors has inspired the creation of novel data access tools in *information visualization* (Fabrikant and Buttenfield, 2001). In the context of visualizing multidimensional data, *spatial metaphors* are particularly relevant, as highlighted in the following paragraphs.

The use of *spatial metaphors* in visualizations is effective, as humans are very familiar with the perception of space and the reasoning behind space (Kuhn, 1996). Spatial orientation and wayfinding are typical daily actions which support humans in understanding space, spatial components, and the relationships between them (Fabrikant and Buttenfield, 2001: 265). Lakoff (1987) describes such recurring experiences and interactions of humans in their physical environment as *image schemata*. Also, *near-far* and *path* are examples of *image schemata* according to Lakoff (1987).

Humans' intuitive understanding of *image schemata* can be capitalized in *information visualization* to depict data in abstract representations. For example, the *near-far schemata* might serve as a metaphorical expression and can be mapped into the *target domain* of *similarity* (Fabrikant, 2000: 67). Therefore, features (e.g., text documents) which are known to be semantically similar are visualized closer to one another in a visualization than features which are not similar. *Multidimensional scaling* (MDS) is one technique used in *information visualization* which explicitly takes the *near-far schemata* into account and seeks to reduce multidimensional data sets to lower dimensional representations (Card, 1996). In Wise et al. (1995), an early approach to depict newspaper articles by applying the MDS technique was presented; newspaper articles which are thematically more similar are depicted closer to one another than newspaper articles which are dissimilar. Further visualization methods, which are based on *spatial metaphor*s, have been adopted

by the *information visualization* community, such as the *self-organizing map* technique which will be discussed in detail in *Subsection 2.2.3*.

The *information visualization* community promotes interdisciplinary research (Card et al., 1991, Keim, 2002). In particular, knowledge from *cognitive science*, *psychology*, and *human-computer interaction* (HCI) has influenced the field of *information visualization*, as these research communities provide extensive research on the cognitive aspects of visualizations as well as on the interaction of users with (interactive) visualizations (Card et al., 1991, Robertson et al., 1991, Card et al., 1999). Additionally, the combination of *information visualization* with the more computational and statistical field of *visual data mining* has proven to be valuable for *knowledge discovery* in large data sets (e.g., Keim, 2002, Shneiderman, 2002).

Drawing from these findings in *information visualization*, the potential for incorporating *spatial metaphors* to study large data archives has been recognized by the GIScience community as well. Kuhn and Frank (1991), Kuhn (1992, 1996), and Kuhn and Blumenthal (1996) studied the meaning of *spatial metaphors* in GIScience from a user interface perspective and introduced the term *spatialization*. Skupin and Buttenfield (1996, 1997), Couclelis (1998), and Fabrikant (2000), adopted the idea of *spatialization* and applied it to large text archives. Fabrikant and Buttenfield (2001) illustrated different spatial frames of reference to formalize and visualize semantic spatialized views, and Skupin and Fabrikant (2003) presented a cartographic research agenda for non-geographic information visualization, highlighting the cartographical and cognitive aspects of spatializations.

Despite these efforts in spatialization research, Fabrikant and Skupin (2005: 668) claimed that until the mid-2000s no systematic approach to formalize the *spatialization framework* based on solid theoretical foundations has been proposed. Therefore, Fabrikant and Skupin (2005) suggest the *spatialization framework* which is depicted in Figure 7.

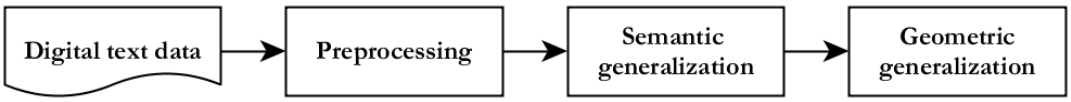


Figure 7: *Spatialization framework* (Fabrikant and Skupin, 2005).

The *spatialization framework* can be applied to various data types (Skupin, 2008). As digital text data is relevant to this thesis, this represents the starting point in Figure 7. Text data is often unstructured or semi-structured; therefore, a preprocessing step is necessary. The preprocessing stage helps to extract relevant information from a raw data source for further processing in subsequent steps of the *spatialization framework*. Applying *geographic information retrieval* methods is a common way to preprocess text data (see *Section 2.1*). The potential outputs of applying *geographic information retrieval* methods to unstructured or semi-structured text archives are lists with occurrences of toponyms and temporal references in the investigated text archives or a *vector space model* of thematic information (see *Subsection 2.1.3*).

Although digital text data is the most common input source for generating spatialized views, the *spatialization framework* can also be applied to structured non-textual data, and therefore, preprocessing steps may not be necessary (Skupin and Fabrikant, 2007). Census data are one example of structured input data for spatialization which was used, for example, by Skupin and Hagelman (2005). Likewise, Koua and Kraak (2005) studied a socio-economic data set and visualized it in a *self-organizing map* (see *Subsection 2.2.3*).

The preprocessed and structured data (e.g., list of toponyms, list of temporal references, vector space model, census data) are then used as an input for *cartographic generalization*. *Cartographic generalization* is a two-step approach that incorporates both the *semantic* and *geometric generalization* in Figure 7 (Fabrikant and Skupin, 2005). The generalization aims at reducing details and complexity for depicting multi-dimensional data in typical 2D maps (Fabrikant and Skupin, 2005: 669-70), and is described in *Subsections 2.2.1* and *2.2.2*. In *Subsection 2.2.3*, applications of the *spatialization framework* which are relevant in the context of this thesis are illustrated.

2.2.1 Semantic generalization

Following preprocessing, *semantic generalization* is applied as illustrated in Figure 7. *Semantic generalization* implies that the attributes of a data set's features are generalized to their essential characteristics (e.g., by applying dimensionality reduction techniques), and *semantic primitives* to express these features with *spatial metaphors* are identified (Fabrikant and Skupin, 2005: 670). According to Lakoff and Johnson (1980), the physical environment (e.g., a geographic landscape) represents the *source*, whereas the conceptual space represents the *target domain* of the metaphorical mapping process which is illustrated in Figure 8.

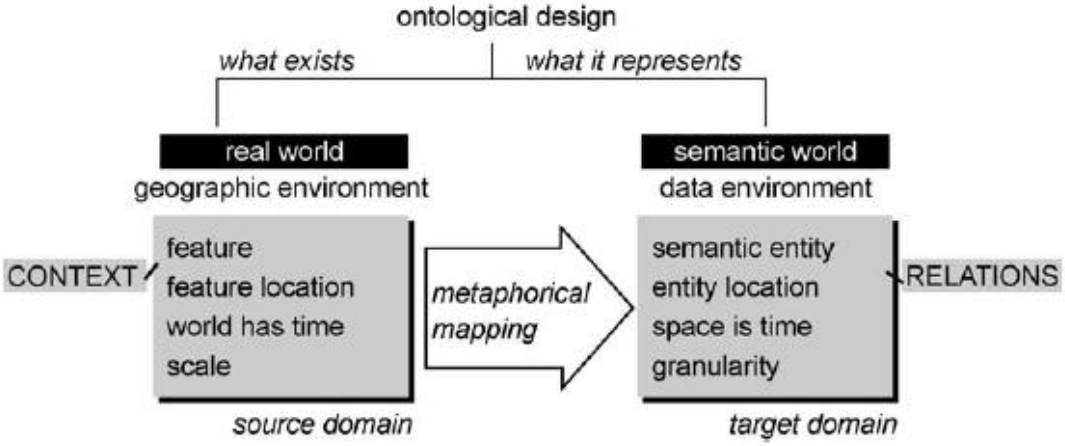


Figure 8: *Semantic generalization* process (Fabrikant and Skupin, 2005: 671).

In Figure 8, *feature*, *feature location*, *time*, and *scale* are part of the physical environment and transferred to *semantic entity*, *entity location*, *space*, and *granularity*, respectively, in the conceptual space. For example, to metaphorically map time (i.e., *world has time* in Figure 8), space (i.e., *space is time* in Figure 8) might be used (Fabrikant and Skupin, 2005: 671). Therefore, more current items might be depicted in the foreground of a spatialization,

whereas less current items might be pushed towards the back of the display (Fabrikant and Skupin, 2005: 671).

Fabrikant and Skupin (2005: 672-73) identify four fundamental representational primitives associated with the physical environment that are applicable to the spatialization of information: *locus*, *trajectory*, *boundary*, and *aggregate*. These four *semantic primitives* are explained in the following list.

- **Locus**. In a spatialized display, information items have a meaningful location or place, metaphorically mapped from the absolute position of features (e.g., a building) in geographic space (Fabrikant and Skupin, 2005: 673). The relative position of an information item in a spatialization is defined by its semantic relationships to other information items. In a MDS, for example, information items (e.g., newspaper articles) which are semantically similar to one another are placed near to one another whereas dissimilar information items are placed distant from one another.

- **Trajectory**. In spatializations, trajectories are a linear entity type and represent semantic relationships between information items that are metaphorically mapped from paths or routes connecting features in geographic space (Fabrikant and Skupin, 2005: 673). Therefore, in the context of newspapers, for example, articles which share the same topic might be connected with a line to depict a strong semantic relationship.

- **Boundary**. In spatializations, boundaries are a linear entity type used to delineate semantic regions that are metaphorically mapped from borders (e.g., country borders) in geographic space (Fabrikant and Skupin, 2005: 673). For example, in the context of newspapers, articles with similar semantic content are clustered in a region and graphically distinguished with a boundary from other regions with articles different in content.

- **Aggregate**. In spatializations, aggregates are an areal entity type and describe semantically similar regions that are metaphorically mapped from regions (e.g., country) in geographic space (Fabrikant and Skupin, 2005: 673). To differentiate *aggregate* from *boundary*, the following example is used: the area of the *United States of America* (USA) is an *aggregate*, whereas the border separating the USA from Canada is a *boundary*. Translated to spatialized displays, this implies that in the context of newspapers, a region with articles similar in content and semantically distinguishable from other regions forms an aggregate.

2.2.2 Geometric generalization

The next step in the spatialization process is the *geometric generalization*, as depicted in Figure 7. *Geometric generalization* is the process of assigning graphic marks or signs to the *semantic primitives* identified during the *semantic generalization* (Fabrikant and Skupin, 2005: 674). *Semantic primitives* are graphically depicted by applying *visual variables* according to

the concept presented by Bertin (1967), which was extended by DiBiase et al. (1992) and MacEachren (1995), and others.

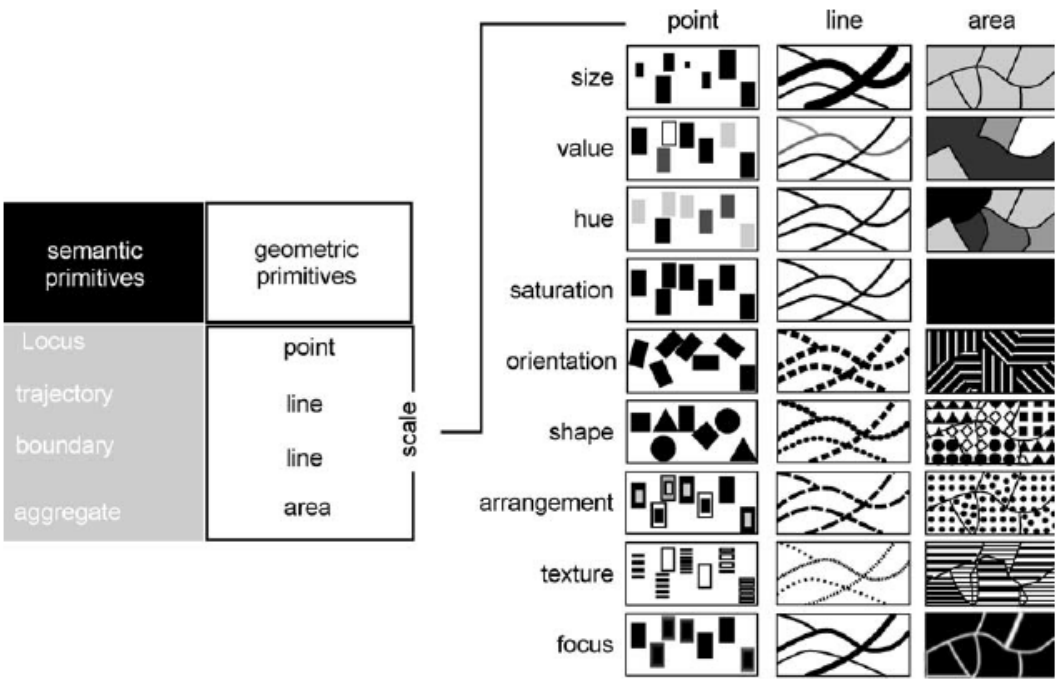


Figure 9: *Semantic primitives*, *geometric primitives*, and *visual variables* (Fabrikant and Skupin, 2005: 674).

The *visual variables* used by Fabrikant and Skupin (2005: 674) for defining the *spatialization framework* are based on work by MacEachren (1995), and are illustrated in Figure 9. *Size*, *value*, *hue*, *saturation*, *orientation*, *shape*, *arrangement*, *texture*, and *focus* are illustrated in Figure 9 as *visual variables* to graphically depict *spatial metaphors* in spatializations. For example, newspaper articles in an information landscape are assigned the semantic primitive *locus*. The geometric primitive *point* is chosen to graphically depict their location in the landscape. Assuming that the articles are classified according to the topic they cover, the visual variable *color hue* might serve to distinguish articles dealing with different topics from one another. Therefore, articles concerning *politics* might be visualized in green, whereas articles concerning *sports* might be visualized in blue.

Geometric generalization is the final step in the *spatialization framework* depicted in Figure 7. In the following section, applications of the *spatialization framework* are presented.

2.2.3 Applying the spatialization framework

The *spatialization framework* is applicable to different data sources and data types. As previously discussed, all data types from unstructured (e.g., raw text) to structured data (e.g., census data) may be used as an input to create spatialized displays.

While the *spatialization framework* is often applied to data without geographic coordinate reference such as raw text data, the framework might be applied to geospatially referenced data as well. Skupin and Hagelman (2005) applied spatialization techniques

to census data to the US county level. However, the *semantic* and *geometric generalization* is based on similarities of the demographic variables in the census data, and the geospatial reference is disregarded in the creation of spatialized displays. Therefore, although the items in the census data (i.e., counties) are geospatially referenced, the similarity values used as an input for the spatialization process are based on non-spatial variables (i.e., demographic variables). Therefore, the creation and layout of spatialized displays (e.g., similarity values) is always based on non-spatial data, even though the items in the source data may have a geographic reference (Skupin and Fabrikant, 2007: 64-67).

Spatialization has promoted different reduction and spatial layout techniques such as the MDS or tree maps. An extensive overview of techniques is provided in Skupin and Fabrikant (2003), whereas we focus on two spatialization techniques which were found to be relevant to this thesis: *network visualizations* and *self-organizing maps*. Network visualizations are considered relevant as they point out relationships and interconnections between data items (Shneiderman, 1996). Relationships are visualized as edges which connect nodes (i.e., data items) in a network. Furthermore, network visualizations highlight hierarchical structures and the centrality of particular data items (Becker et al., 1995). Therefore, they might serve to depict spatio-temporal and thematic relationships in large text archives and thus might be useful to answer *Research Question 2* of this thesis (see *Section 1.3*). Applying network visualizations to geographic information in large text archives has been demonstrated, for example, by Salvini (2012), which is further detailed in this section.

The second spatialization technique we found to be relevant to this research project are *self-organizing maps*. Similar to the network visualization approach, *self-organizing maps* depict multidimensional input data in a two-dimensional visualization (Skupin and Agarwal, 2008). However, the relationships between data items are visualized on a map instead of in a network (Skupin and Agarwal, 2008). In contrast to network spatializations, connections between data items are not visually highlighted with an edge. Instead, the location of data items in a map illustrate their semantic relatedness: the higher the semantic similarity of two data items, the closer they are placed to each other in the *self-organizing map*. Therefore, *self-organizing maps* are particularly helpful in identifying and analyzing clusters of semantically related data items. In contrast to network visualizations, *self-organizing maps* are particularly useful for depicting very large data sets (i.e., thousands or millions of data items) (Skupin and Agarwal, 2008). Furthermore, they can process bimodal input data such as a *term-document matrix* or a *topic-document matrix* (see *Subsection 2.1.3*), which is illustrated with an example in this subsection.

Considering *network visualizations* and *self-organizing maps* in the context of this thesis further complies with Moretti's (2005) suggestion to use *graphs*, *maps*, and *trees* in order to visualize large text data archives in the humanities: the *network visualizations* are a typical *graph drawing* technique, the *self-organizing maps* visualize data on a *map*. In the following paragraphs, we first detail the *network visualization* approach and then focus on *self-organizing maps*.

Network visualization

The visualization of network-display spatializations is based on the *distance-similarity metaphor* which is empirically evaluated in Fabrikant et al. (2004). The *distance-similarity metaphor* in spatializations suggests that items which are semantically similar are placed closer to one another on the network and connected with an edge (i.e., a line), compared to items which are semantically dissimilar. This metaphor is based on the *first law of geography*, which states: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970: 236). Montello et al. (2003) extended this law to spatializations, as spatializations are experienced by people as if they were geographic spaces (i.e., people *believe* closer things are more similar) and coined this the *first law of cognitive geography*. For point spatializations (e.g., MDS), the *first law of cognitive geography* has been empirically evaluated by Montello et al. (2003). The results support the *first law of cognitive geography* because straight-line distance between points has been understood by study participants as a metaphor for dissimilarity (Montello et al., 2003). However, in network spatializations there are conflicting notions of distance: *straight-line metric* (i.e., direct distance), *network metric* (i.e., distance along the network), and *topological proximity* (i.e., number of nodes between source and destination node). Fabrikant et al. (2004) empirically tested the *first law of cognitive geography* in network spatializations and concluded that the *distance-similarity metaphor* operates, but the *network metric* is most influential with regard to the perception of distance compared to other notions of distance.

In the following paragraphs, an example of a network spatialization presented by Salvini and Fabrikant (2016) is briefly discussed. This example was chosen as it deals with the computation and depiction of spatial relationships in network visualizations, and the information was retrieved from *Wikipedia*[27] article texts. Therefore, similar to the approach in this thesis, Salvini and Fabrikant (2016) seek to analyze and visualize geographic information in a large online data archive and focused on text data in applying the *spatialization framework*. Methodological details are covered in Salvini and Fabrikant (2016). The approach we applied to visualize network spatializations in this thesis is shown in *Chapter 4 – Methods*.

Salvini and Fabrikant (2016) investigated the English version of *Wikipedia*, a massive, crowd-sourced, freely accessible multimedia online database in order to study the multirelational world city network. An XML file of the English version of *Wikipedia* was accessed (i.e., semi-structured data), and important tags used for creating the network were stored in MySQL database tables. 95 cities were identified as world cities, and relationships between these cities were computed following the principle illustrated in Figure 10.

In Figure 10, the *Wikipedia* article *Barack Obama* is a *shared article*, which implies that at least the *Wikipedia* articles of two world cities are referenced as hyperlinks (i.e., *New York* and *Chicago* in the *shared article Barack Obama*). This is interpreted by Salvini and Fabrikant (2016: 233-34) as a relationship between the two cities. The more existing

[27] English version of Wikipedia: https://en.wikipedia.org/ (accessed June 2016)

shared articles in which two cities are referenced, the stronger the relationship between the two cities. Details regarding the calculation and weighting of relationships are illustrated in Salvini and Fabrikant (2016: 233-34). This approach was inspired by Hecht and Raubal (2008), who also used the hyperlink structure of *Wikipedia* to assess semantic relationships between lexically expressed entities.

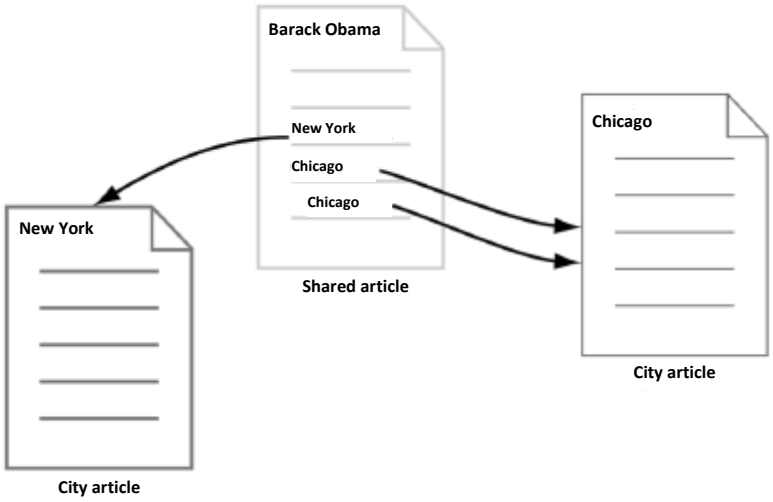


Figure 10: Measuring intercity relations based on *Wikipedia* hyperlink structure (Salvini and Fabrikant, 2016: 233).

In the next step, *shared articles* were clustered based on their category tags on *Wikipedia*. This clustering was based on a *topic modeling* approach used to describe the category tags of the articles quantitatively in an *article-topic matrix*, which was followed by a subsequent clustering of the articles following the *community detection approach* introduced by Blondel et al. (2008). These steps are detailed in Salvini and Fabrikant (2016). As a result, themes have been assigned to each of the shared articles (e.g., *politics* for *Barack Obama*). Eight themes (e.g., *economy/technology*, *politics*, *sports*) were identified, and for each of these themes a world city network was modeled.

Following these preprocessing steps, *semantic generalization* for depicting the relationships in a network spatialization was performed: world cities were interpreted as *loci*, and relationships between them as *trajectories*. During *geometric generalization*, the geometric primitive *point* was assigned to the semantic primitive *locus* and *line* to *trajectory*, respectively. In addition, the visual variables *size*, *color hue*, and *color value* were applied to depict the *centrality* of cities, the membership of cities to *world regions*, and the *strength of relationship*, respectively (see the *economy/technology* world city network in Figure 11).

In order to create the network visualization in Figure 11, all *shared articles* classified as *economy/technology* were considered, and the *pathfinder network scaling* algorithm (Dearholt and Schvaneveldt, 1990) was employed in order to simplify the network and extract the structurally most important relationships between the world cities. To depict these important relationships visually, the *graph embedder* (GEM) algorithm was applied (Frick et al., 1995) which places strongly connected world cities close to one another, thereby

complying with the *distance-similarity metaphor* in network visualizations (Fabrikant et al., 2004). Furthermore, *visual variables* were employed: different *color hues* represent the membership of cities to geographic regions. The *size* of the nodes (i.e., points) stands for the centrality of the cities in the world network. Thus, the more and the stronger connected to other world cities, the higher the centrality of a city and therefore the larger the node in the network. The *color value* of the edges (i.e., lines) depicts the strength of the relationship between two cities. The stronger a relationship (i.e., the more two cities are referenced in *shared articles*), the darker the color of an edge. In addition, the edges of strongly connected cities are visualized thicker than edges between weakly connected cities, and thus the visual variable *size* is used in combination with *color value* to highlight *relationship strength*. The variable *functional centrality* in Figure 11 is not explained here, though interested users are referred to Salvini and Fabrikant (2016: 240).

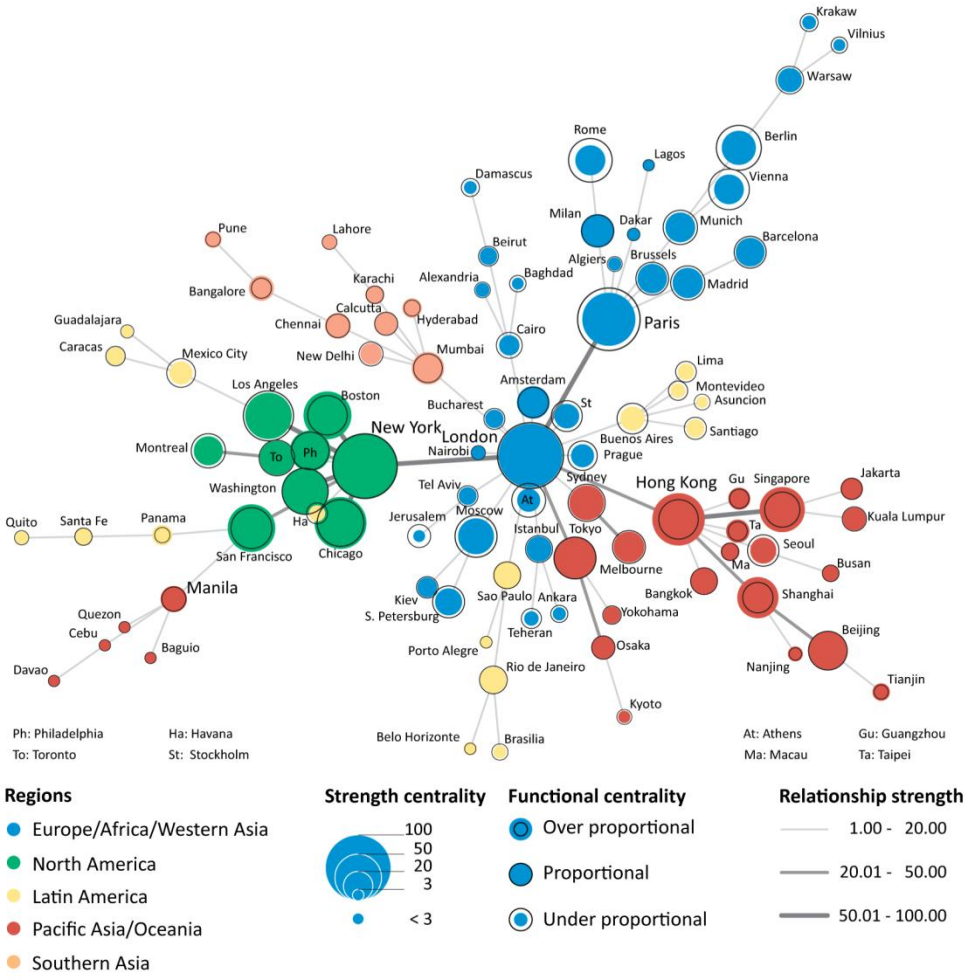


Figure 11: Economy/technology network with 95 world cities (modified from Salvini, 2012: 164).

The presented application of the *spatialization framework* to crowdsourced data illustrates how *spatial metaphors* can be used to visualize text data and to study spatial relationships, structures, and the hierarchical organization of spatial data in a text archive. New research questions might evolve by presenting the achieved results to urban

geographers, and further research projects in *global city* research could be developed. An extensive discussion of the results and the methodological details can be found in Salvini (2012). A similar approach is described in Fabrikant and Salvini (2011), who analyzed the semantic content of titles and abstracts submitted to the *International Cartographic Conference* (ICC) 1999-2009, then visualized the result in a network spatialization. In Skupin (2014), another network spatialization is presented which depicts a co-citation network of an influential researcher in order to understand the researcher's influence on the evolvement of specific topics in a scientific domain.

Incorporating the temporal dimension in network visualizations has been addressed by the *dynamic network visualization* community (e.g., Bender-deMoll and McFarland, 2006, Bach et al., 2014, Muelder et al., 2014, Hadlak et al., 2015). Particularly relevant in the context of this thesis is a debate in this community regarding *mental map preservation*. *Mental map preservation* refers to the preservation of node locations over time in dynamic network visualizations. For example, if a *Node A* is located at the top left and *Node B* in the top right of a network in the beginning of the study period, then these nodes must also be located at the top left and top right, respectively, in the network of the second period of time if the mental map layout should be preserved. Several user studies were conducted to test if *mental map preservation* supports users in performing tasks in dynamic network visualizations (i.e., *error rate* and *response time*). However, the results are contradicting, as Purchase and Samra (2008), Saffrey and Purchase (2008), Archambault et al. (2011), and Archambault and Purchase (2012) state that *mental map preservation* has no influence on the performance (i.e., *error rate* and *response time*), whereas Archambault and Purchase (2013) argue that better performance results are achieved if the mental map layout is preserved.

In this section, we illustrated network visualizations as a useful means to analyze the structure and interconnection of geographic information in large text archives. In the following section, we introduce a second spatialization technique that is relevant to our research project: *self-organizing maps*.

Self-organizing map

The *self-organizing map* (SOM) is an example of an *artificial neural network* (ANN) technique (Kohonen, 2001), and could be interpreted as a combination of *clustering* and *dimensionality reduction* techniques (Skupin and Agarwal, 2008). In short, SOMs spatially cluster objects which share the same characteristics (Skupin and Agarwal, 2008). This procedure also complies with the *first law of cognitive geography* (Montello et al., 2003), and was empirically tested in Fabrikant et al. (2006).

In the following paragraphs, an example of a SOM approach presented by Skupin and de Jongh (2005) is briefly discussed. This example was chosen because it deals with the computation and depiction of semantic relationships in a SOM and the information analyzed in the example was retrieved from a large digital text archive. Therefore, similar to our approach, Skupin and de Jongh (2005) sought to analyze and visualize semantic structures and connections in a large text archive by applying the *spatialization framework*.

Different to the aforementioned network visualization technique, the SOM of Skupin and de Jongh (2005) depicts connections based on the thematic attributes (i.e., *term-document matrix*) of input data instead of connections based on co-occurrences of toponyms in text documents as per the network visualization. We focus on the application of the *spatialization framework*, whereas methodological details will not be covered. Interested readers are referred to Kohonen (1990, 2001). Further information regarding how SOMs were incorporated into this project is presented in *Chapter 4 – Methods*.

Skupin and de Jongh (2005) investigated the proceedings of the *International Cartographic Conference* (ICC) 2001 in Beijing, China and 2003 in Durban, South Africa [28]. The original text documents were transformed from Word and PDF format to annotated XML files (i.e., tags for *identifier*, *title*, *author name and address*, *keywords*, *abstract*, *full text*). The XML files were parsed into a relational database. After stop word removal, *Porter stemming* was executed (Porter, 1980) in order to strip suffixes, reduce singular/plural forms to a common root, and employ further manipulations. Then, terms occurring in the *titles*, *abstracts*, *full texts*, and *keywords* of the papers were transformed into a *term-document matrix* based on a *vector space model*. The fields in the matrix were filled with term counts for each document, similar to Figure 5 (left).

After these preprocessing steps, *semantic generalization* for depicting the papers in a SOM was applied: papers were interpreted as *loci*. In the *geometric generalization* stage, the geometric primitive *point* was assigned to the semantic primitive *locus*. The SOM with individual ICC papers is depicted in Figure 12.

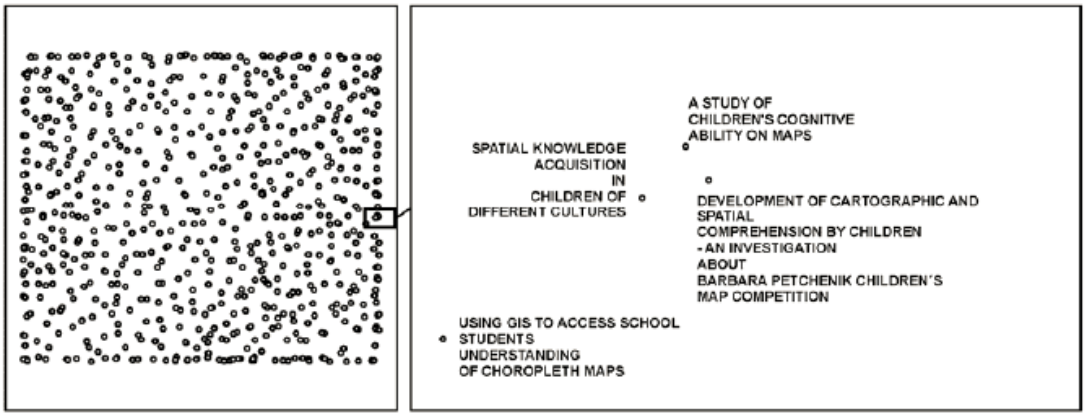


Figure 12: Spatialization of single ICC conference papers (Skupin and de Jongh, 2005).

In Figure 12 (left), the entire SOM consisting of all 708 ICC papers considered in this study is depicted. In Figure 12 (right), a zoom in on the region framed by a black rectangle in the SOM is illustrated. In addition, the points are labeled in the zoomed-in view with the respective paper title.

[28] International Cartographic Conferences: http://icaci.org/icc/ (accessed June 2016)

The SOM was created by applying a two-step neural network training approach (Skupin and de Jongh, 2005). Firstly, global structures were assessed and the different regions in the SOM were assigned different topics. For example, in Figure 12, the region marked for the zoom-in could be about *GIS and children*, as the presented paper titles on the right indicate. Secondly, the structures on a more local level were optimized. Then, ICC papers, visualized as points in Figure 12, are placed in the region of the SOM which semantically best represents the content of the paper. In other words, papers which are close together in the resulting SOM are assumed to be semantically similar (i.e., similar terms are used in the papers). The neural network training stages are covered in detail in Skupin and Agarwal (2008).

A further possibility for visualizing data in a SOM is to cluster the single ICC papers semantically and visualize the clusters on top of the SOM. For the *semantic generalization*, the clusters are interpreted as *aggregates*. During *geometric generalization*, the geometric primitive *area* was assigned to the semantic primitive *aggregate*. The SOM with semantic clusters on top is illustrated in Figure 13.

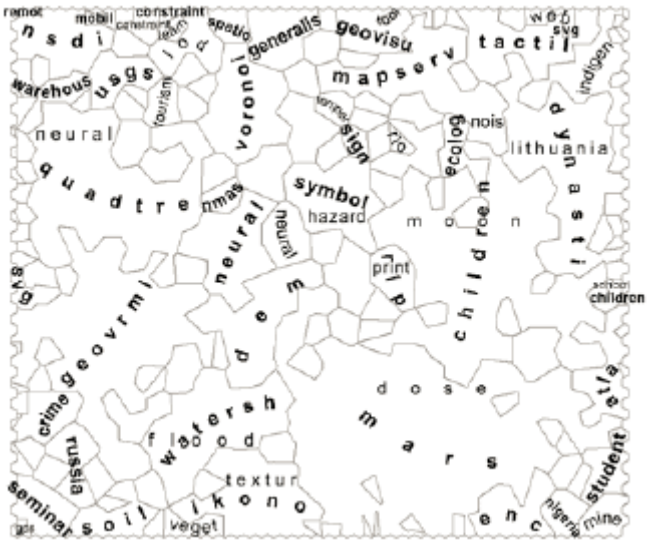


Figure 13: *Self-organizing map* with labeled clusters (Skupin and de Jongh, 2005).

In order to cluster the ICC papers visualized in Figure 12, the *k-means* algorithm (MacQueen, 1967) was applied to the *term-document vectors* of the papers. The *k-means* algorithm groups elements such that elements within a group are very similar compared to elements outside the group. This is completed by minimizing within-group and maximizing between-group variance. In other words, elements (i.e., papers) which use similar terms, and therefore possess similar *term-document vectors* are grouped in the same *k-means* group, whereas elements with dissimilar *term-document vectors* are assigned to different *k-means* groups. After applying the *k-means* algorithm, each ICC paper in the SOM belongs to a *k-means* group. Mathematical details of the *k-means* algorithm are reported in MacQueen (1967).

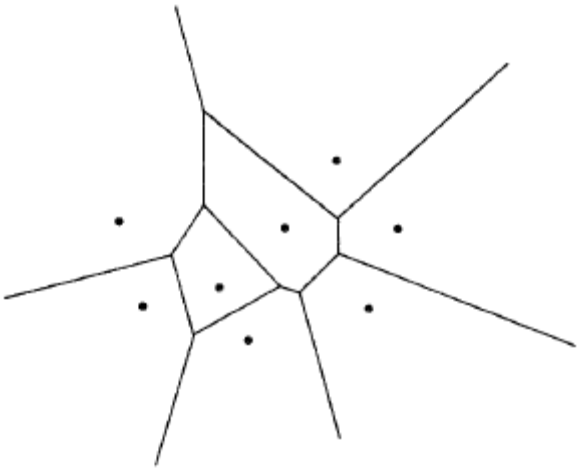


Figure 14: Thiessen polygons for eight points (Aurenhammer, 1991: 347).

In the next step, *Voronoi tessellation* was employed (Okabe et al., 2000). Thus, for each point in the SOM, a region is defined in which the distance to each point compared to the other points in the SOM is shortest. On the borders of these regions, the polygon boundaries (i.e., *Thiessen polygons*) are drawn as depicted in Figure 14 for an eight-point solution. If on both sides of the resulting *Thiessen polygons* the respective points belong to the same group, borders were dissolved. In other words, if neighboring papers in the SOM (see Figure 13) belong to the same *k-means* group, the border between them is dissolved. For the resulting clusters, labels were assigned by employing the *tf-idf* method (see *Subsection 2.1.3*).

The presented application of the *spatialization framework* to conference papers illustrates another example of using *spatial metaphors* to visualize text data. People interested in cartography and in particular in the ICC might study the SOM and derive interesting insights regarding the coverage of topics and semantic relationships between topics in the information landscape. In addition, the *first law of cognitive geography*, which holds true for the SOM (Fabrikant et al., 2006), can be capitalized by interested people to judge similarity of papers in the SOM based on the distance between them (Montello et al., 2003). In Skupin and de Jongh (2005), the SOM regarding the ICC is further discussed. A similar approach was chosen by Skupin (2002), who visualized 2,220 abstracts submitted to the *Annual Meeting of the Association of American Geographers* (AAG) in Honolulu, Hawaii (1999) in a hierarchical cluster SOM. Further related text-based SOM approaches are reported in Skupin (2009) and in Skupin et al. (2013) who visualized knowledge domain and the topical structure of medical sciences, respectively. Steiger et al. (2016) proposed a geographic, hierarchical SOM to analyze the geospatial, temporal, and semantic characteristics of tweets, similar to Hagenauer and Helbich (2013) who analyzed and visualized spatio-temporal data from different data sets by using hierarchical SOMs. Andrienko et al. (2010a) also worked with spatio-temporal data and proposed space-in-time and time-in-space SOMs to explore spatio-temporal patterns. Wang et al. (2013) provided another example of applying the SOM approach in geography; they visualized meteorological data with a temporal dimension using a SOM.

Summary

- Spatialization uses *spatial metaphors* to depict multidimensional data in lower dimensional visualizations.
- After input data are preprocessed and structured, they are generalized to *semantic primitives* (i.e., *locus*, *trajectory*, *boundary*, *aggregate*) and depicted using *geometric primitives* (i.e., *point*, *line*, *area*) and *visual variables* (e.g., *size*).

We discovered that *network visualizations* and *self-organizing maps* might be relevant to answering *Research Question 2* of this thesis, as they both support analyzing and visualizing spatio-temporal and thematic structures and interconnections in large digital text archives. In the next step, we aim at incorporating these spatialized displays in interactive and exploratory web interfaces in order to provide target users access to the spatio-temporal and thematic information in text data. In addition to the *distant reading* idea of Moretti (2005), to abstract and reduce information in text archives in the humanities and depict them in visual displays (as illustrated in this section), we are interested to allow target users direct access to the raw data, and thus combine the *distant reading* concept with *close reading*. *Geovisual analytics* provides a methodological framework for the development of combined *distant* and close *reading* exploratory web interfaces, involving target users in the entire interface design and evaluation process. Involving target users helps to assess and incorporate specific requirements of target users in the interface design and evaluation process which is detailed in the following section.

2.3 Geovisual analytics

Thomas and Cook (2005, 2006) introduced the field of *visual analytics* and summarized it as "…the science of analytical reasoning facilitated by interactive visual interfaces. People use *visual analytics* tools and techniques to synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data; detect the expected and discover the unexpected; provide timely defensible, and understandable assessments; and communicate assessment effectively for action" (Thomas and Cook, 2006: 10). Furthermore, Thomas and Cook (2006) promote the development of *visual analytics* tools for collaboration, the creation of a taxonomy for interaction techniques, the systematic evaluation of *visual analytics* tools, and the provision of a robust *privacy and security* infrastructure. Therefore, *visual analytics* combines *visualization*, *human factors*, and *data analysis* while drawing from different research fields such as *information visualization and analytics*, *data management* and *data mining*, *knowledge discovery*, and *cognitive, perceptual*, and *interaction science*, as visualized in Figure 15 (Keim et al., 2006).

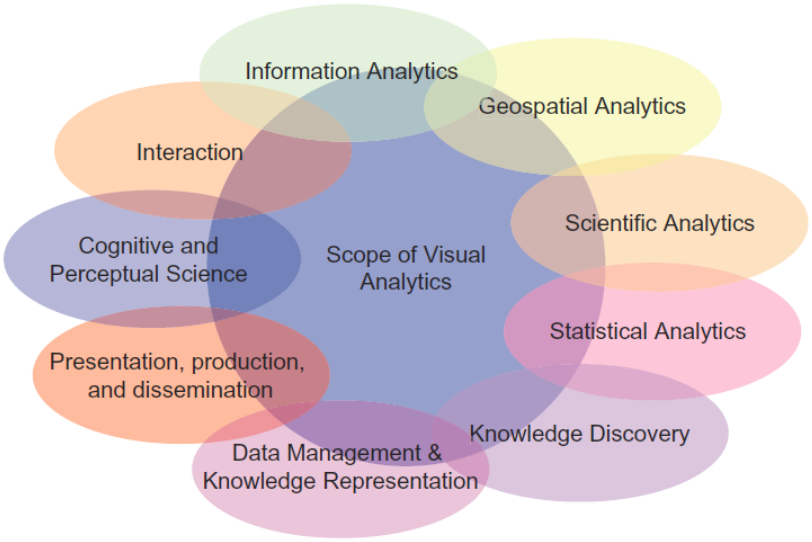


Figure 15: *Visual analytics* as a highly interdisciplinary research field (Keim et al., 2006).

Highlighting *geospatial analytics* in the scope of *visual analytics* (see Figure 15) has initiated the formation of the *geovisual analytics* (geoVA) community. GeoVA is a sub-area of *visual analytics* with a special focus on space and time (Andrienko et al., 2007: 841). Andrienko et al. (2010b: 1579) emphasize the need for appropriate geoVA tools that are easily accessible and usable not only for highly qualified specialists, but for all people. Kraak (2008) interprets the development of geoVA from a cartographic perspective as depicted in Figure 16.

Kraak (2008) links the development of *cartography* to *computer cartography*, *geovisualization*, and *geoVA* to the steady increase in the amount and diversity of data, as illustrated in Figure 16. *Cartography* was influenced strongly by *art*, *design*, *geography*, and *surveying* (Kraak, 2008). In the mid-1980s, advances in computer and communication technologies had a huge impact on *cartography* resulted in the establishment of *computer cartography* (Kraak, 2008). This form of technologically driven *cartography* incorporated knowledge from fields such as *exploratory data analysis*, *geographical information systems*, and *scientific visualization* (Kraak, 2008). In the mid-1990s, the creation of the *information visualization* community (see *Section 2.2*) and the emerging field of *GIScience* induced the formation of *geovisualization*. *Geovisualization* primarily focuses on new technological possibilities to dynamically and interactively depict complex data, support group work, and promote human-centered approaches (Kraak and MacEachren, 2005). In the mid-2000s, *visual analytics* initiated the formation of geoVA.

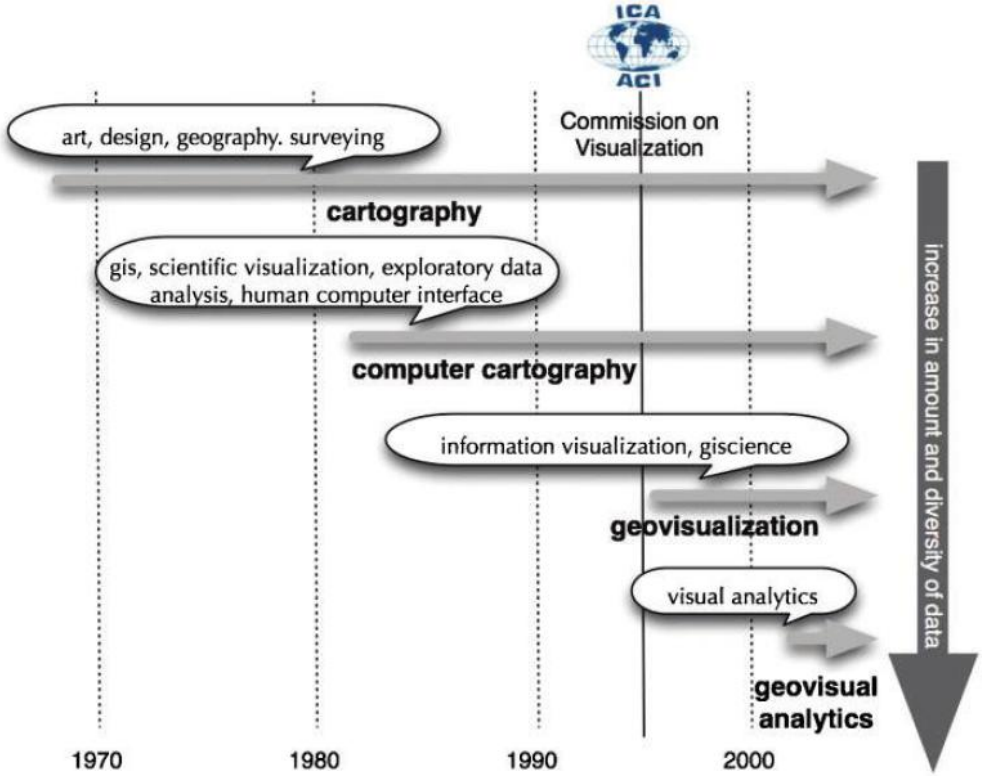


Figure 16: Timeline from *cartography* to *geovisual analytics* (Kraak, 2008: 163).

In the following subsections, we focus on spatio-temporal information as well as user-centered design and evaluation approaches in geoVA, as a large portion of this thesis is based on these central aspects of geoVA. Additionally, some examples of recent geoVA applications are introduced.

2.3.1 Space and time in geovisual analytics

Space

Spatial data possess certain characteristics which require special treatment in geoVA and GIScience in general (e.g., Anselin, 1989, Goodchild, 2002, Haining, 2009). Five characteristics of spatial data are particularly relevant for working with and analyzing spatial data: *spatial dependence*, *spatial heterogeneity*, *large volume*, *uncertainty*, and *scalability* (Anselin, 1989, Goodchild, 2002).

Spatial dependence refers to the tendency that geographic data exhibit *spatial autocorrelation*. This implies that characteristics of features at proximal locations are likely to be positively or negatively correlated (Goodchild, 2002). This is the subject of the *first law of geography* from Tobler (1970). The existence of *spatial dependence* has severe implications for spatial data analysis as it violates the primary assumption of hypothesis testing, which expects samples to be drawn independently from a population (Goodchild, 2002). *Moran's I* (Moran, 1950) or *Geary's c* (Geary, 1954, Jeffers, 1973) are statistical methods used to test *spatial autocorrelation* and measure the correlation of adjacent observations in space. Both measures test global *spatial autocorrelation* patterns. Anselin (1995) and Getis

and Ord (1992) present approaches to decompose these global indicators into the contributions of single observations, and thus support the detection of local spatial clusters and spatial outliers. If *spatial autocorrelation* is identified, common data analysis methods such as *regression analysis* must be adopted. Anselin et al. (2010) describe a spatial regression method which accounts for *spatial autocorrelation* by incorporating the values of the dependent variable in close spatial proximity as an additional independent variable in the regression model in order to decrease the autocorrelation of regression model errors (i.e., *spatial lag/ spatial error model*). Readers interested in spatial regression are referred to Anselin (2003).

Spatial heterogeneity refers to the tendency that conditions vary from one spatial location to another (Goodchild, 2002). Therefore, the results of an analysis at a location differ from the same spatial analysis at another location, which is called spatial *non-stationarity*, a state which does not allow for the possibility for generalizations of processes for geographic regions with varying conditions (Goodchild, 2002). For example, to account for *spatial heterogeneity* in *regression analysis*, *geographically weighted regression* (GWR), might be applied (e.g., Leung et al., 2000, Fotheringham et al., 2002, Fotheringham, 2009, Wheeler, 2014). GWR calibrates a multiple regression model that allows various regression coefficients to exist at different locations in space (Brunsdon et al., 1996). Thus, regression coefficients are locally estimated using a subset of observations surrounding a location, whereas spatially close observations have a higher influence on the local model parameters than observations that are spatially distant (Wheeler, 2014). Readers interested in GWR are referred to Fotheringham et al. (2002).

Goodchild (2002) identifies *large volume* and *uncertainty* as further typical characteristics of spatial data. For example, the text of a medium-sized novel is about one *megabyte* (MB) in size, whereas a remotely sensed image occupies hundreds of MB. *Uncertainty* is related to errors due to measurement inaccuracy, vague definitions of classes, errors introduced during processing, and other reasons.

A final important characteristic of spatial data is the existence of spatial phenomena at different spatial *scales*. The *scale* is defined by the extent of the frame a spatial data set is viewed and aggregated at. For example, if comparing different regions in the world, a spatial analyst might choose *continents* (i.e., large scale) or single *countries* (i.e., small scale) as *scale*, depending on the goal of the research project. The choice of an appropriate *scale* to analyze data is very important, as in an extreme case, opposite relationships might be uncovered if viewed at an inappropriate spatial *scale* (Andrienko et al., 2010b: 1584). Furthermore, if aggregating spatial data to larger units (e.g., from *countries* to *continents*), spatial analysts must be aware of the *modifiable areal unit problem*, as the analysis results might depend on the strategy used to aggregate data (Openshaw, 1983). Not only the size of aggregates (e.g., a *continent* as an aggregate for many *countries*), but also the delineation of aggregated regions (i.e., how boundaries are drawn) influences spatial data analysis (Wong, 2009). Therefore, Andrienko et al. (2010b: 1584-85) suggest testing the *sensitivity* of any findings to the method of spatial aggregation.

The concept of *space*, but also the concept of *time* is important to geoVA. Similar to *space*, *time* has special characteristics which need to be considered in geoVA. These characteristics of temporal information are covered in the following section.

Time

Time has an inherent hierarchical structure and thus operates at various temporal *scales*: *seconds*, *minutes*, *hours*, *days*, *weeks*, etc. (Andrienko et al., 2010b: 1582). Similar to the *spatial scale*, the choice of the *temporal scale* and aggregation strategy (e.g., aggregate 60 consecutive seconds to 1 minute) influences spatio-temporal analyses in GIScience, as shown by Laube and Purves (2011), who studied the influence of choosing different temporal scales on the analysis of movement data. In addition, the choice of an appropriate *temporal primitive* to model time influences the results of temporal data analysis. Time can either be modeled as *time points* (i.e., one instant in time) or *time intervals* (i.e., temporal primitive with an extent), depending on the goals of the analysis (Andrienko et al., 2010b: 1582-84). These primitives are organized in different *calendar systems* (e.g., *Gregorian* calendar, *Julian* calendar), and *natural cycles*, and *re-occurrences* of time (e.g., seasons) (Andrienko et al., 2010b: 1582).

Similar to the spatial dimension, *dependence* and *autocorrelation* must be considered in temporal data analysis because the feature attributes of consecutive time steps are likely to be positively or negatively correlated. Therefore, applying standard hypothesis testing, which assumes independence among observations, is often not possible (Andrienko et al., 2010b: 1583). However, *temporal dependence* could be used to estimate missing values in a time series through *interpolation* or *extrapolation* (Andrienko et al., 2010b: 1583). One possibility to test *temporal dependence* is to employ the *runs test*, which evaluates the randomness of the distribution of elements in a sequence based on the *median value* of the sample (Wald and Wolfowitz, 1940, Bradley, 1960). For example, if all data values in the first half of a sequence are below the *median*, and data values in the second half are above the *median*, the test indicates that the elements in a sequence are not randomly distributed, and the analyst must assume that a temporal trend exists in the data.

As we have illustrated in this section, certain characteristics of spatial and temporal information must be considered for geoVA applications and tools because space and time are the most important information dimensions in geoVA. Furthermore, as stated at the beginning of this section, providing exploratory and interactive visual interfaces which facilitate information seekers to gain new insights regarding spatio-temporal and thematic information and structures in a data set is a primary goal of geoVA. To reach this goal, applying a user-centered design and evaluation approach is suggested by the geoVA community, which is covered in the following subsection.

2.3.2 User-centered design and evaluation in geovisual analytics

The development of a systematic approach to design, develop, evaluate, and empirically test geoVA applications with users is central to geoVA as it promotes the synergistic relationship between humans and machines to generate knowledge and new insights

from complex spatio-temporal data sets (Andrienko et al., 2010b, Roth and MacEachren, 2016). The following overview is based on Roth et al. (2015), who study systematic user-centered design and evaluation approaches in geoVA. The approach by Roth et al. (2015) is considered relevant to this thesis, as it illustrates a typical geoVA design and evaluation approach and because the interface concept of our geoVA interfaces is similar to the concept of Roth et al. (2015) as illustrated in this and the next subsection. Additionally, Lewis and Rieman's (1993) work on task-centered user interface design is considered. Lewis and Rieman (1993) incorporate an evaluation stage without users, which we found to be a useful and relevant extension to our approach. This is further illustrated in the proceeding text.

In geoVA, *interface success* is not only based on good programming, debugging, and adhering to cartographic principles (e.g., *visual variables*, *symbolization*), but it also strongly depends on essential interface characteristics such as how easy the interface is to learn and to use, if all requested functionalities are integrated effectively, and if the target group is able to work with the interface (Roth et al., 2015). Therefore, designing a successful geoVA user interface requires a deep understanding of potential users and their needs as well as an iterative design process with multiple evaluation and revision stages (Roth et al., 2015: 263). Roth et al. (2015: 264-67) define *interface success* in a triangle of three components: *user*, *utility*, and *usability*. These components are incorporated in an iterative process, starting with the *user*, as depicted in Figure 17. The target group must first be defined. Then, the requirements and tasks of target users must be assessed and specified (Lewis and Rieman, 1993: 11-19). Requirements and tasks might be assessed by observing target users (e.g., *interaction logs*) in their life or work environment (DeWalt and DeWalt, 2002), interviewing target users (Rosson and Carroll, 2002a), or by conducting a *focus group research*. In a *focus group*, a few target users discuss very preliminary ideas and suggest possible adaptations (Rubin and Chisnell, 2008: 17).

Figure 17: *Interface success* based on an iterative *user-utility-usability* design process (Roth et al., 2015: 267).

Once the requirements of target users are assessed and specified, *utility* is the next step in the triangle in Figure 17. *Utility* is defined as the usefulness of an application to users in order to accomplish users' objectives (Grinstein et al., 2003). Therefore, geoVA designers define a functionality baseline for the geoVA application based on the results

of the requirements and tasks analysis (Roth et al., 2015: 266). In the next step, designers implement the initial prototypical versions of an application. The prototypical versions are compared to one another, and usable interface designs, meeting the required *utility* baseline, are identified (Roth et al., 2015: 266). This step is the *usability* part in Figure 17. *Usability* is defined as the ease of using an interface in order to accomplish a set of users' objectives (Grinstein et al., 2003). The prototypical version is then presented to target users, which completes the *user-utility-usability* triangle in Figure 17. Target users interact with the geoVA application and evaluate the *usability* of the prototypical version as well as provide feedback regarding possible *utility* improvements for a subsequent version of the application (Roth et al., 2015: 266). Once this feedback is received, designers are prompted to conduct another *user-utility-usability* run (Roth et al., 2015: 266). Many of these runs can be necessary for a geoVA application to be ready for full release.

Contrary to the model presented in Figure 17, Lewis and Rieman (1993) suggest the substitution of the *user* by the *designer* in an early *user-utility-usability* run. Lewis and Rieman (1993: 41) argue that a designer must evaluate the design first prior to testing it with users, because an interface presented to users should be as free of problems as possible. One formal method for testing a design without users is the *cognitive walkthrough* method (e.g., Lewis and Rieman, 1993, Wharton et al., 1994, Lewis and Wharton, 1997, Mahatody et al., 2010). In a *cognitive walkthrough*, the interface designer attempts to imagine users' actions with the interface (e.g., pressing a button, clicking on a hyperlink) for a set of representative tasks (Lewis and Rieman, 1993: 46). The *designer* attempts to tell a believable story if and how users might solve a task. If the designer is unable to tell a believable story, a user may not take the correct action to solve a task, and a potential problem with the interface is discovered (Lewis and Rieman, 1993: 46). The *designer*'s judgments in a *cognitive walkthrough* are based on her or his knowledge of target users gained at an earlier stage of the user-centered design process (Lewis and Rieman, 1993: 48). Potential problems identified during the *cognitive walkthrough* evaluation are eliminated. Only then, the geoVA application is presented to *users* in a subsequent *user-utility-usability* run.

In the existing literature, many methods for empirically testing and evaluating *utility* and *usability* with users are described. Here, only two of these methods that are relevant to this thesis and to geoVA in general are briefly outlined: *surveys* and *think aloud studies*. Readers interested in further information regarding *utility* and *usability* evaluation are referred to Landauer (1995) for a discussion about *usefulness*, *usability*, and *productivity*, to Rosson and Carrol (2002b) for advice regarding *usability engineering*, to Rubin and Chisnell (2008) for a generic overview about *usability testing*, and to Tullis and Albert (2013) for recommendations on how to collect, analyze, and present *usability metrics*. Readers interested in *usability* metrics assessed using *eye-movement analysis* are referred to Çöltekin et al. (2009).

Surveys and *think aloud studies* are common user-based *utility* and *usability* methods. *Surveys* may be conducted as *formative* or *summative* studies (Tullis and Albert, 2013: 45-47). *Formative* studies have an iterative character, and *utility* and *usability* may be assessed

several times in the early and intermediate stages of a project (Roth et al., 2015: 268). During a *formative* study, problems are identified, the interface is revised, and several rounds of evaluations might be executed (Tullis and Albert, 2013: 46). For example, Roth et al. (2015: 279) conducted a *formative* online survey in order to assess *usability* metrics such as *learnability*, *efficiency*, *memorability*, *error rates/severity*, and *subjective satisfaction* in a geoVA interface. Furthermore, *utility* was tested by assessing the *effectiveness* of different use case scenarios, including the *novelty* and the *comprehensiveness* of included functionality in their geoVA interface (Roth et al., 2015: 279-80). In contrast, *summative* studies are conducted at the end of a project to test if the *utility* and *usability* goals of the project are met (Tullis and Albert, 2013: 46-47). For example, Roth et al. (2015: 279-80) conducted a *summative* online survey after the aforementioned *formative* online survey and used the same questions to compare the results of the non-revised (i.e., before the *formative* study) and the revised version (i.e., after the *formative* study) of their geoVA interface.

Conducting *think aloud studies* is another possibility for the assessment of the *utility* and *usability* of geoVA applications. During a *think aloud study*, participants are asked to perform a task and comment on their thoughts while working on it (Lewis and Rieman, 1993: 83). Comments can encompass questions that arise, things participants read, or simply describe what participants are doing (Lewis and Rieman, 1993: 83). The moderator of the study is not allowed to answer any questions during the study, and can only intervene if participants stop commenting on their thoughts for a certain amount of time (Boren and Ramey, 2000: 263). *Think aloud sessions* usually are audio- or video-taped, and *utility* and *usability* issues are identified and categorized by the designer according to how *important* and how *difficult* they would be to repair (Lewis and Rieman, 1993: 85-86). The *importance* judgment is based on the costs of the problem to potential users (e.g., in time), and a prediction regarding the number of users that would be affected by the same problem (Lewis and Rieman, 1993: 86). Issues which are ranked *important* and issues which are ranked *easy* to solve are repaired by the designer (Lewis and Rieman, 1993: 86).

So far, we have illustrated central aspects of geoVA: the characteristics of *space* and *time*, and a *user-centered design and evaluation* approach used to develop geoVA applications. In the following subsection, we illustrate the application of the geoVA framework in previous works.

2.3.3 Applying geovisual analytics

In recent years, many geoVA applications have been developed. In this subsection, some of them, which we considered relevant in the context of this thesis, are discussed.

The geoVA application by Roth et al. (2015), which was used to illustrate user-centered design and evaluation in *geovisual analytics* in the previous subsection, is particularly relevant to our thesis. Roth et al. (2015) depict criminal activity on an interactive map in a geoVA interface. On one hand, the geoVA interface incorporates several spatial and temporal filtering functionalities and thus allows for *distant* and *close reading* of criminal

activity data. On the other hand, Roth et al. (2015) applied an iterative *user-utility-usability* approach to design and evaluate the geoVA interface, which is relevant to this thesis as we aim to involve target users in the design and evaluate of geoVA interfaces (see *Section 1.3*). A detailed description of the *interaction primitives* (e.g., filter, search, zoom) used in the criminal activity interface developed by Roth et al. (2015), and an empirical interaction study based on these *interaction primitives* are found in Roth et al. (2014) and Roth and MacEachren (2016).

In other related research projects, geoVA methods are combined with GIR to retrieve spatio-temporal and thematic information from text data which is relevant to our project as we consider GIR methods to answer *Research Question 1*, and geoVA approaches to answer *Research Question 3* (see *Section 1.3*). For example, Tomaszewski (2008) presented a geoVA approach to produce *geo-historical* context from *Really Simple Syndication* (RSS)[29] feeds. GIR methods were applied to geocode locations mentioned on websites which were referenced in RSS feeds. Furthermore, GIR methods were employed to locate important *concepts* (i.e., relevant terms) on these websites. Based on user queries, relevant locations and concepts, as well as connections between them (based on *co-occurrences* of locations and concepts on websites) are visualized in a *multiple linked views* visualization using *Google Earth*[30] as an interface base map. Additionally, temporal filtering based on the time stamp of the RSS feeds (i.e., the publishing date) allow for the retrieval of information relevant to a specified time frame. Similarly to Tomaszewski (2008), Robinson et al. (2016) presented a geoVA interface based on RSS feeds. However, Robinson et al. (2016) focused on identifying temporal patterns of events by analyzing spatial, temporal, and thematic content in the text data on the websites that the RSS feeds refer to. Events on these websites are thematically coded (e.g., *politics*, *economy*), and if event types occur significantly often temporally close to one another (e.g., event type *threaten* often occurs before event type *confront*), an event pattern is identified. These event patterns are visualized in a *multiple linked views* interface with various *filtering* options (e.g., *timeline view* for temporal filtering, *map view* for spatial filtering). If users select an interesting event pattern and apply the available *filtering* options, all relevant websites are presented in a list to them. In addition, Robinson et al. (2016) provided an evaluation approach to test the *utility* and *usability* of their geoVA interface.

Other examples in geoVA incorporate spatialized displays in interactive geoVA interfaces. This is relevant to our thesis as it addresses *Research Questions 2* and *3*, since we discovered the *spatialization framework* to be relevant in depicting spatio-temporal and thematic structures and interconnections (see *Section 2.2*). Additionally, geoVA suggests techniques to depict the spatialized displays in exploratory and interactive web interfaces. For example, Luo et al. (2014) presented the *GeoSocialApp*, which combines traditional *network analysis* methods with a *map* view (i.e., choropleth map) in order to analyze geo-social relationships in the international trade network. In the *network view*,

[29] Find a definition of *Really Simple Syndication* (RSS) here: http://www.webopedia.com/TERM/R/RSS.html (accessed June 2016)
[30] Google Earth: https://www.google.com/earth (accessed June 2016)

the relationships between countries based on the amount of trade between them are depicted. Countries linked with strong trade connections are grouped and separated from countries to which these countries only have weak trade connections. In the *map view*, the spatial distribution of the *gross domestic product* (GDP)[31] is visible. The *network* and the *map view* are dynamically linked to each other. The exploration of the combined *network* and *map view* allows for the generation of new insights regarding complex interactions between spatial and social relationships. Fabrikant et al. (2015) illustrated a geoVA approach to interactively explore census data attributes in a simple web-based application, combining a SOM and a *map view* (i.e., choropleth map). First, a user enters a desired spatial (e.g., *region*) and temporal unit (e.g., *year dates*) in the web tool. Then, the corresponding SOM is automatically created and displayed to the user. The resulting SOM represents a socio-demographic landscape, meaning that features (e.g., *regions*) with similar census data attributes are visualized in close proximity to one another, and features with dissimilar attributes are visualized distant from one another. An interactive *map view* is linked to the SOM which allows for the analysis of single census data attributes in geographic space.

Summary

- *Geovisual analytics* originates from *visual analytics*, and facilitates analytical reasoning in large spatio-temporal and thematic data sets by providing a framework to create and evaluate interactive visual user interfaces.
- While applying *geovisual analytics* to spatio-temporal data, researchers have to consider special characteristics of space and time such as *spatial* and *temporal dependence* and *autocorrelation*.
- *Geovisual analytics* promotes the use of an iterative user-centered design and evaluation approach, including *utility* and *usability* evaluations and involving target users in the early stages of an interface design process.

We discovered that geoVA is relevant to answering *Research Question 3* of this thesis, as it provides a framework to combine multidimensional data in exploratory and interactive web interfaces. In addition, geoVA suggests interactive web interfaces with coupled *distant* and *close reading* functionalities (e.g., accessible by zooming and filtering in the previously presented geoVA interfaces), as proposed by Jockers (2013). Furthermore, geoVA suggests user-centered design and evaluation methods for the development of interactive applications. Involving target users in an iterative design and evaluation process early on is particularly relevant to this thesis for the purpose of determining the needs and requirements of target users in the humanities. Combining computing technologies and the humanities, and the combination of GIScience and the humanities is covered in the following section.

[31] Find a definition of *Gross Domestic Product* (GDP) here: http://www.investopedia.com/terms/g/gdp.asp (accessed June 2016)

2.4 Digital humanities

Digital humanities (DH) have experienced a huge interest in recent years, and an increasing number of researchers contribute to this relatively new research field (Gold, 2012, Kaplan, 2015). We introduce DH as we consider a case study in the humanities in this thesis (see *Chapter 3 – Data*) and contribute our approach to ongoing research in DH, which seeks computing methods to get new insight into the humanities. First, briefly we introduce DH in *Subsection 2.4.1*, and then illustrate combined humanities and GIScience approaches with a particular focus on text data in *Subsection 2.4.2*, as this is particularly relevant to this thesis.

2.4.1 Defining the digital humanities

Defining the field of DH and delineating the boundary of this research field has been extensively and controversially discussed in the literature (e.g., Svensson, 2009, Ramsay, 2011, Terras, 2011, Kirschenbaum, 2012). Historically the field of DH emerged from the common ground of computing and linguistics and was termed *humanities computing* (Svensson, 2009). In the early 2000s, Schreibman et al. (2004) edited a book regarding this emerging research field entitled *A Companion to Digital Humanities* in order to place less emphasis on technological aspects and opened the field to humanists in general (Fitzpatrick, 2012: 13). From then onward, the term DH has become popularized, and, in the subsequent years, the DH community has grown steadily (Fitzpatrick, 2012: 13). Not only literary study experts, but also scholars from *history*, *musicology*, and *media studies* as well as other related fields in the humanities joined the DH community to benefit from incorporating modern computing technologies into their field of research (Fitzpatrick, 2012: 13). As a consequence of this opening to a wide range of humanities disciplines, very diverse research approaches have been adopted that have produced significant tension within the DH community (Fitzpatrick, 2012: 13). In particular, a controversial discussion has begun regarding whether or not DH should only be related to *making* (i.e., using digital technologies in studying traditional humanities) or if it should also be expanded to include *interpreting* (i.e., using methods of contemporary humanities in studying digital objects) (Fitzpatrick, 2012: 13-14). Although some DH scholars insisted on the central role of the *making* component in DH (e.g., Ramsay, 2011), DH research seems to have bridged the *making* and *interpreting* components in productive ways in recent years (Fitzpatrick, 2012: 14, Kaplan, 2015).

Instead of developing a rigid definition of DH and discussing who is part of it and who is not, Spiro (2012: 16-17) suggested the development of a flexible statement of *shared values* of the DH community. Spiro (2012: 23-30) proposes the following core values: *openness* (i.e., exchange of ideas, open content and software, and transparency), *collaboration*, *collegiality and connectedness*, *diversity*, and *experimentation*. In a similar vein, Svensson (2012) advises neither to define DH tightly nor draw boundaries, but sees DH as a trading zone and meeting place, which is highlighted by the following quote.

> *"…the digital humanities needs to support and allow multiple modes of engagement between the humanities and the digital in order to touch at the heart of the disciplines, maximize points of interaction, tackle large research and methodology challenges, and facilitate deep integration between thinking and making. This perspective would seem to be compatible with the digital humanities as a trading zone and a meeting place. (…). Whether mostly physical or mostly digital, they can help channel dispersed resources, technologies, and intellectual energy. Furthermore, deep integration of toolmaking and interpretative perspectives requires very different kinds of competencies and work to happen in the same space. It could also be argued that there is value to unexpected meetings in creative environments in terms of expanding the digital humanities."*
>
> Svensson (2012: 46)

To summarize, DH is still ill-defined (Kaplan, 2015). Generally speaking, one might view DH as an interdisciplinary research endeavor which seeks to study new methods of expanding the humanities with computing technologies and fosters interdisciplinary research collaborations between humanities experts and experts in modern digital technologies.

2.4.2 Digital humanities and GIScience

At the nexus between geography and the humanities, a few research sub-fields have evolved. For example, *GeoHumanities* seeks to connect the concepts of space and place with the humanities (e.g., Dear, 2015, Hawkins et al., 2015). Likewise, *spatial humanities* attempts to connect the concept of space with the humanities, but focuses particularly on the introduction of GIS technologies to the humanities (e.g., Bodenhammer et al., 2010, Bodenhammer et al., 2013). Similarly to *spatial humanities*, but with a special focus on history, the *historical GIS* field has emerged (e.g., Gregory and Ell, 2007, Gregory and Healey, 2007). All of these research areas are not easily distinguishable from one another; therefore, we will treat them as one sub-field of DH. Within this spatial DH sub-field, we will focus on research projects at the nexus of GIScience and the humanities which deal with text data in this subsection as this topic is particularly relevant to this thesis.

Text data has been central to many humanities disciplines (e.g., history, literature) long before digitalization. With recent advances in information and communication technology, new opportunities to analyze and depict text data have been created. Franco Moretti, a literary scholar (see *Section 1.2*), has influenced literary studies in DH considerably. He works on gaining new insights from text data in the humanities by applying computational methods and published an influential work on the analysis and visualization of a collection of novels in *graphs, maps, and trees* in the early years of DH research, which is used at different sections of this thesis to motivate the applied research approach (Moretti, 2005). Moretti's (2005, 2013) *distant reading* approach implies that literary scholars might not have to read every single work in a text collection, but could rather use *information analysis* and *visualization* techniques (e.g., *graph visualization*) to gain new insights into the overall interconnections of a studied data set. By highlighting

the potential of using *maps* and geographical concepts in literary studies to gain new insight into the humanities, Moretti reinforced the collaboration of the DH community with GIScience experts (Moretti, 2005).

A further DH scholar mentioned in previous sections of this thesis to motivate our project is Matthew L. Jockers. Jockers (2013) highlights the potential for studying an entire corpus by applying computational methods which he calls *macroanalysis* (i.e., *distant reading*), analogous to the sub-field of *macroeconomics* in economics, which investigates the economic behavior of an entire economy. The counterpart in economics is *microeconomics*, which examines the behavior of individual consumers and businesses. Analogous, Jockers (2013) calls the study of individual texts *microanalysis* (i.e., *close reading*). Jockers (2013) highlights the potential of *macroanalysis* as a new approach to supplement *microanalysis*, rather than substituting it, which is exemplified by the following statement: "The result of macroscopic investigation is contextualization on an unprecedented scale. The underlying assumption is that by exploring the literary record writ large, we will better understand the context in which individual texts exist and thereby better understand those individual texts. This approach offers specific insights into literary historical questions (…)" (Jockers, 2013: 27). Among other approaches, Jockers (2013: 118-53) applies *probabilistic topic modeling* (see *Subsection 2.1.3*) to 3,346 works of American, British, and Irish fiction to understand the thematic structure of the studied works. Furthermore, Jockers (2013: 154-68) uses the output of the *probabilistic topic modeling* and additional stylistic attributes of the works to depict them in *spatialized network visualizations*.

Applying GIR techniques to automatically retrieve spatial, temporal, and thematic information from large unstructured digital text archives has been investigated by many researchers in DH (e.g., Berzak et al., 2011, Clifford et al., 2016, Donaldson et al., 2016). Existing gazetteers and GIR systems to retrieve information from digital text archives have been optimized for historical text document collections (e.g., Grover et al., 2010, Southall et al., 2011, Alex et al., 2015, Gregory et al., 2016). Using spatialized displays to depict relationships is as well common in DH, but is mostly limited to the visualization of social networks (e.g., Ciula et al., 2008, Bingenheimer et al., 2011, Tóth, 2013), or the depiction of relationships between texts of a corpus (e.g., Weingart and Jorgensen, 2013, Reiter et al., 2014). Highlighting spatial or spatio-temporal relationships is often done by superimposing lines on a base map (e.g., Barker et al., 2010, Evans and Jasnow, 2014). Further projects situated at the nexus between GIScience and DH are presented in Jänicke et al. (2015). In the following paragraphs, we focus on two projects which are relevant to this thesis because they combine various GIScience techniques and apply them to digital text collections in the humanities.

In Gregory and Hardie (2011) and Murrieta-Flores et al. (2015), a combined corpus linguistics and GIScience approach is presented that supports the exploration of historical text data. GIR methods are applied to annotate place names in historical text documents. Then, *concordances* are studied (Murrieta-Flores et al., 2015: 302). Thus, words that occur near place names in text documents are extracted and define the thematic context of places (Murrieta-Flores et al., 2015: 298). For example, in the sentence

"*London* is a leading *financial* center", *London* is identified as a place name and since *financial* co-occurs in the same sentence as *London* within a short textual distance (i.e., four words) *financial* is assumed to describe *London* thematically. This contextual information is then depicted on a map. Therefore, if *finances* is defined as a search term, all locations which co-occur very often with financial terms in text documents might be highlighted on the map (Gregory and Hardie, 2011).

A similar approach to Gregory and Hardie (2011) and Murrieta-Flores et al. (2015) regarding the retrieval and combination of spatial and thematic information in text documents is applied in Hinrichs et al. (2015): *commodity trading* is explored by analyzing four large historical text document collections employing GIR and information visualization approaches. *Locations*, *commodities* (e.g., *sugar*, *coal*), and *temporal information* are semi-automatically retrieved and stored in a relational database (Hinrichs et al., 2015). A relationship between a location and a commodity is established if they co-occur in the same sentence (Hinrichs et al., 2015). For example, in the sentence "They were bringing *coal* to *London*", a *location-commodity relationship* between *coal* and *London* is assumed. These relationships are visually mapped in an interactive user interface (Hinrichs et al., 2015). A user might be interested in *coal* and enters the query term *coal* in the interface. Then, all locations which co-occur with *coal* in the four historical text document collections are highlighted on a map and the user may select interesting locations and access original documents in which the commodities and locations are mentioned (i.e., coupled *distant* and *close reading* functionalities). Furthermore, temporal filtering options are available. Interested readers are referred to Hinrichs et al. (2015) and to the project website[32].

These aforementioned works at the nexus of DH and GIScience represent only a small sample of combined research endeavors. Nevertheless, they illustrate a high interest of the DH community in GIScience methods, and thus demonstrate an opportunity for GIScience to contribute to DH and apply well-established methods in GIScience to humanities data sets. An ongoing debate in DH regarding *aesthetics* and the *usability* of interfaces (e.g., Kirschenbaum, 2004, Drucker, 2011a, Drucker, 2011b, Gibbs and Owens, 2012) highlights another point of potential collaboration and exchange between GIScience (i.e., *geovisual analytics*) and DH.

Summary

- *Digital humanities* is an interdisciplinary field that seeks to introduce a wide range of computing technologies to the humanities.
- Combining humanities with GIScience approaches is a growing subfield of *digital humanities* that attracts scholars to initiate interdisciplinary research collaborations.

[32] Trading Consequences project website: http://tradingconsequences.blogs.edina.ac.uk/ (accessed June 2016)

In this section, we have illustrated the emerging field of DH which attempts to link the humanities to fields which are more focused on computing technologies. In the context of our research project, the link between DH and GIScience is particularly relevant as it demonstrates the general need for and growing interest in incorporating GIScience and geographic methods in the humanities. As illustrated in this section, analyzing text data in the humanities from a GIScience perspective has been investigated by several DH scholars in recent years. Our research project aims at contributing to such research endeavors.

The research gap that we identified by studying related work in all relevant fields of this research project is introduced in the following section.

2.5 Research gap

In the previous sections of this chapter we illustrated related work which is relevant to this thesis and identified the following research gap.

Research gap

A comprehensive approach a) to automatically retrieve spatial, temporal, and thematic information from unstructured or semi-structured text data in the humanities, b) to transform the retrieved multidimensional data into lower dimensional spatialized displays, c) to develop interactive user interfaces of spatialized displays to allow information seekers to explore spatio-temporal and thematic information, structures, and interconnections in large unstructured and semi-structured text archives in the humanities, and gain new insights into the humanities from a spatio-temporal and thematic point of view, following a user-centered design and evaluation approach, is currently absent from the existing literature.

GIR provides methods to bridge the first part of the research gap. We aim at developing a combination of existing and appropriate GIR methods to automatically retrieve information about space, time, and theme from a typical text archive in the humanities (see *Chapter 3 – Data*) and adapt existing approaches, if necessary. Although GIR suggests a multitude of methods for the retrieval of multidimensional and spatial data from unstructured and semi-structured data archives, to the best of the author's knowledge, no comprehensive approach has been presented thus far for the automatic retrieval of spatial, temporal, and thematic information from unstructured and semi-structured text data in the humanities, such that spatio-temporal structures and interconnections can be generated for further processing and spatialized display. Therefore, we introduced the following research question in *Chapter 1 – Introduction*.

- **Research Question 1**. How can information about space, time, and theme be automatically retrieved from unstructured and semi-structured text archives in the humanities so that hidden structures and relationships can be uncovered in the data?

The second part of the research gap considers the transformation and reorganization of multidimensional data retrieved from the text archive into lower dimensional spatialized displays, and the subsequent integration of spatialized displays in interactive user interfaces that involve target users in the entire design and evaluation process. The *spatialization framework* has been applied to many text data sources employing a multitude of spatialization methods (e.g., *network visualization* and *SOM*). *Geovisual analytics* propagates the development of user interfaces to gain new insights and generate new hypotheses regarding large spatio-temporal and thematic data sets by following a user-centered design and evaluation approach. However, to the best of the author's knowledge, no comprehensive approach to create spatializations based on spatio-temporal and thematic data retrieved from unstructured or semi-structured text data in the humanities, and to incorporate them in a *geovisual analytics* framework, involving target users in the entire design and evaluation process has been presented to date. The *spatialization* and the *geovisual analytics* part of the research gap are addressed by *Research Questions 2* and *3* of this thesis.

- **Research Question 2**. How can we spatialize uncovered spatio-temporal and thematic structures and interconnections extracted from unstructured and semi-structured text archives in the humanities?

- **Research Question 3**. How can we make spatialized information about space, time, and theme from unstructured and semi-structured text archives available to information seekers in the humanities to support sense-making and the generation of new insights about these text archives?

Bridging the research gap contributes to the emerging field of *digital humanities*, which seeks new approaches to incorporate novel digital methods in the humanities. Furthermore, we consider Moretti's (2005) *distant reading* concept and Jockers' (2013) suggestion to combine *distant* and *close reading* functionalities as we aim at providing interested information seekers in the humanities with an interactive interface which incorporates both concepts and should help target users to gain new insights and develop new research hypotheses about spatio-temporal and thematic information, structures, and interconnections in the humanities.

In this section, we have presented the research gap we attempt to bridge with our research project. In order to illustrate our research approach, we decided to conduct a case study in the field of history, which is presented in the following chapter.

3 Data

In the previous chapters, we illustrated our idea to analyze large digital text archives in the humanities from a GIScience perspective. The main aim is to provide information seekers in the humanities with coupled *distant* and *close reading* functionalities to identify and analyze spatial, temporal, and thematic information, structures, and interconnections in the text archives. We chose the *Historical Dictionary of Switzerland* (HDS)[33] as a case study to apply our research approach for several reasons: the HDS represents a semi-structured and multilingual digital text archive in the humanities. More than 110,000 articles describe the history on the territory of today's Switzerland in all languages of the country (i.e., *German*, *French*, *Italian*, and *Rhaeto-Romanic*) and covers all time periods and eras since the *Paleolithic*, which started approximately 1.5 million years ago within the territory of today's Switzerland, until today (Le Tensorer, 2015). The articles are grouped into *thematic contributions* (e.g., historical phenomena and terms, institutions, companies), *geographical entities* (e.g., municipalities, mountains, rivers) *biographies*, and articles about important *families* in Swiss history (Morosoli, 2000: 10-11). Thus, the HDS articles do not only contain a great deal of implicit and explicit spatial and temporal information, but also a wealth of thematic information due to the wide variety of topics which are covered, as outlined in *Subsection 5.1.3*. Thus far, spatio-temporal and thematic information in the HDS have not been retrieved from all articles or systematically analyzed and depicted, and only limited querying options (i.e., title or full text query, article category filtering, alphabetical order) are available in the current online version. In Moretti's (2005) and Jockers' (2013) words, this implies that only *close reading* is possible and no *distant reading* functionalities have been implemented yet. Due to these reasons, the HDS fits very well as a case study for this research project. Since the HDS was chosen, we defined target users in this project as historians who are interested in new media types and methods in history, people who are interested in *digital humanities*, and those who are interested in interactive interfaces with which to explore the humanities in general. Digital versions of the HDS in *German*, *French*, and *Italian* have been provided for this project courtesy of the HDS *editorial office*.

In the following subsections, information regarding the creation of the HDS is provided, and the current version of the dictionary is described in detail. Furthermore,

[33] Homepage of the Historical Dictionary of Switzerland (German version): http://www.hls-dhs-dss.ch/d/home (accessed June 2016)

the planned future development of the HDS is presented, and the use of this data source in the context of this thesis is critically discussed. Sources that are referenced in the following subsections are primarily in German. However, a short description of the HDS is available in English on the HDS website[34].

3.1 History of the Historical Dictionary of Switzerland

Switzerland has a long history of encyclopedias dating back to the *medieval times*, which is discussed by Jorio (2004). For the development and creation of the HDS, the following two encyclopedias were of particular importance: the *Allgemeines Helvetisches, Eydgenössisches oder Schweitzerisches Lexicon* (= *universal Helvetian, federal or Swiss lexicon*, author's translation), and the *Historisch-biographisches Lexikon der Schweiz* (= *historical-biographical lexicon of Switzerland*, author's translation) (Jorio, 2004). The *Allgemeines Helvetisches, Eydgenössisches oder Schweitzerisches Lexicon* (AHESL) was written by Johann Jakob Leu (1689-1768) and published between 1747 and 1765, and consists of 20 volumes with historical, geographical, genealogical, and biographical information about the *Old Swiss Confederacy* (Jorio, 2004: 107-08). From 1786 to 1795, Hans Jakob Holzhalb (1720-1807) added six supplementary volumes to the AHESL (Jorio, 2004: 108). The AHESL was published in German only, and was the standard reference work on Swiss history at that time (Jorio, 1998). In the 19th century, no efforts to create a new encyclopedia comparable to the AHESL were made and the *Historisch-biographisches Lexikon der Schweiz* (HBLS), which was initiated by Victor Attinger (1856-1927) and published between 1921 and 1934, was the first attempt to substitute the AHESL in a new encyclopedic work (Jorio, 1998). The HBLS consists of seven original volumes as well as one supplementary volume, and all articles were published in German and French (Jorio, 1998). The HBLS has remained the most important reference work on Swiss history until the end of the 20th century (Jorio, 1998).

In the second part of the 20th century, requests by politicians and scientists for a new and revised version of the HBLS began to emerge, but only in 1987 did the Swiss federal parliament approve the HDS project and grant financial resources (Morosoli, 2000: 10). In 1988, the HDS foundation was established, and a detailed concept was elaborated with an article key word list (Morosoli, 2000: 10). The concept of the HDS was ambitious, as more than 36,000 articles about Swiss history were planned to appear in *German*, *French*, and *Italian* (Morosoli, 2000: 9). In addition, two *Rhaeto-Romanic* versions containing together about 3,100 articles[35], particularly regarding the history of the *Rhaeto-Romans* and *Grisons*, was planned (Jorio, 2000: 202-03). *Grisons* is a *canton* (i.e., administrative unit one stage below country level) in the southeastern part of Switzerland in which the *Rhaeto-Romanic* culture and language is present. Following the idea of the HBLS, the HDS planned to cover Swiss history comprehensively, though the articles should be based on recent findings in Swiss history research, with an emphasis

[34] English description of the Historical Dictionary of Switzerland: http://www.hls-dhs-dss.ch/english.php (accessed June 2016)
[35] Number taken from: http://www.hls-dhs-dss.ch/redac/downloads/pressetext2010_LIR_de.pdf (accessed June 2016)

placed on the 19th and 20th centuries (Morosoli, 2000: 10). 20% of HDS content was intended to encompass the 19th and another 20% the 20th century (Morosoli, 2000: 10). The *medieval period* and *early modern age* were both assigned 25%. The remaining 10% were reserved for the time period between the *Paleolithic* and the *medieval period* (Morosoli, 2000: 10). Furthermore, topics only scarcely described in the HBLS, but of high interest to current history researchers (e.g., *social* and *economic history*), were planned to be covered in greater detail (Morosoli, 2000: 10).

In the process of writing and publishing articles, many different actors were involved, as described by Morosoli (2000: 11-16) and summarized briefly as follows: the *editorial office* defines topics to be covered in the HDS and contracts *authors*. *Scientific advisers* support the *editorial office* in selecting relevant articles, suggest *authors* who may write the respective articles, and assess the quality of articles. Each article category has structural guidelines which are specified by the *editorial office*. For *biographies*, very strict rules regarding the content and the sequence of information in the articles must be followed in order to describe historically important people as comprehensively, but also briefly as possible. Similarly, for *geographical entities*, strict structural rules must be applied. In contrast, the rules for *families* and *thematic contributions* are less strict due to the heterogeneity of contents contained in these articles. For some articles, *authors* cannot follow all rules strictly, as, for example, the *date of death* of some historically important people is not known and this information is missing in the respective *biographies*. After an *author* has finished writing an article, the article is proofread by one of the *scientific advisers* and sent back to the *author* for potential corrections. Then, the article is sent to the *editorial office*, which checks for formal correctness and evaluates other criteria in order to judge if the article fits well into the dictionary. Possible adaptations are discussed with the *author* and the article is then ready for publication. In the final step, *translators* translate the articles from the original language (e.g., *German*) to the other two languages (e.g., *French* and *Italian*), which includes a further cycle of proofreading and correcting. In total, more than 2,500 *authors*, about 100 *scientific advisers*, 75 *translators*, and 40 people working at the *editorial office* (i.e., *main office* in Bern, *other offices* in Bellinzona and Chur) have contributed to the HDS[36].

The HDS should serve humanities experts as well as those in the general public that are interested in history as a research resource, and it was thought to be published in a series of printed books at first (Jorio, 2000: 198). However, in response to digital advances in the 1990s, the HDS decided to make the finished articles freely available online, prior to the first printed volume being published (Jorio, 2000: 198-99). Thus, in September 1998, the *e-Historical Dictionary of Switzerland* (e-HDS) went online containing approximately 8,000 articles in German, French, and Italian (Jorio, 2000: 199). The e-HDS was extended and continuously updated with new articles until it reached more than 36,000 articles in each language version by the beginning of 2014 (Jorio, 2014, September 13). The first of thirteen printed volumes of the HDS was published in 2002 and each year a new volume was presented, such that by the end of 2014, all articles were published in printed form (Jorio, 2014, September 13). The two printed volumes of the

[36] Numbers taken from: http://www.hls-dhs-dss.ch/d/mitarbeiter/zentralredaktion (accessed June 2016)

Rhaeto-Romanic HDS version were published in 2010 and 2012, respectively, and all articles are freely available online in the e-LIR[37]. Different to the printed versions, no graphics, maps, and pictures are available in the e-HDS and the e-LIR versions yet (Jorio, 2000: 200).

3.2 The e-Historical Dictionary of Switzerland

The most recent version of the e-HDS in *German*, *French*, and *Italian* was provided by the HDS for this research project as XML[38] files on January 30, 2015. The files contain all article texts, including some metadata. In this project, we only considered the German version of the HDS. We decided to use the German version, as the author of this thesis as well as all participants in the user studies (see *Subsection 4.3.1*) have German as their native language. Providing participants with information and interfaces in German helped minimize problems and errors resulting from misunderstanding or misinterpreting the language. Due to choosing the German version, all numbers and figures presented in this subsection are based on this language version. The articles in different language versions are not word-by-word identical, but equal that the sense of the original articles being conveyed. Therefore, the content of articles in different language versions is very similar and thus all reported findings in this section are similar for the other language versions.

In Table 1, the number of articles in the e-HDS is divided into article categories and the corresponding percentages. In the final column, the average article length in words is listed.

Table 1: HDS article categories.

Article category	Total	Percent	Length
Biographies	25,202	69.6	128
Geographical entities	5,350	14.8	422
Thematic contributions	3,067	8.5	625
Families	2,569	7.1	183
Total	36,188	100	218

Biographies is the most prominent article category in the HDS, accounting for nearly 70% of all articles, whereas *families* accounts for 7.1%. Articles in these two categories are the shortest, with an average article length of 128 and 183 words, respectively. The remaining 23.3% of the HDS articles are divided into *geographical entities* (14.8%) and *thematic contributions* (8.5%). Articles in these categories are longer than the *biographies* and *families* articles, with an average article length of 422 and 625 words, respectively. The average length of all 36,188 articles is 218 words.

[37] Homepage of the Historical Dictionary of Switzerland (Rhaeto-Romanic version): http://www.e-lir.ch/ (accessed June 2016)
[38] Introduction to XML: http://www.w3schools.com/xml/ (accessed June 2016)

If the total amount of words in the HDS is considered, *geographical entities* and *thematic contributions* articles account for 28.6% and 24.4%, respectively, whereas *biographies* and *families* contribute 41% and 6%, respectively. This outlines that, although, *biographies* are the shortest in average length, as shown in Table 1, they contribute the greatest amount of words to the HDS in total, which is due to the high overall frequency of *biographies* in the HDS (i.e., 25,202 articles).

Searching for articles in the HDS is possible either by entering title or full text query terms in the e-HDS. Furthermore, there is an option to assess articles by their alphabetical order (similar to the printed version). The search interface of the e-HDS in its current version (i.e., June 2016) is depicted in Figure 18.

Figure 18: e-HDS full text search interface (retrieved from HDS, 2016b).

In Figure 18, the *full text search* (= *Volltextsuche*) with *advanced filtering options* is selected and *Zürich* (i.e., name of a Swiss canton and the largest city of Switzerland) has been entered as a search query. A user might select between *title and text* (= *Titel und Text*) and *title only* (= *nur Titel*) search, between the languages *German* (= *Deutsch*), *French* (=*Français*), and *Italian* (= *Italiano*), and might further *constrain the search* (= *Suche einschränken auf*) to specific article categories by checking one or more than one of the boxes labeled *geographical entities* (= *Ortsartikel*), *thematic contributions* (= *Sachartikel*), *biographies* (= *Biographien*), and *families* (= *Familienartikel*). If no box is selected, all article categories are considered. At the bottom of Figure 18, the title of the article ranked highest according to the HDS retrieval system (i.e., *Zürich (Kanton)*), including the first lines of the article text, are displayed. The query term (i.e., *Zürich*) is highlighted in red in the article text. Additionally, the number of *characters* in the article (= *Zeichen*) and the *author*

(= *Autor*) of the respective article is displayed. By clicking on the title (i.e., *Zürich (Kanton)*), a new window containing the respective article pops up.

In Figure 19, the first part of the article *Zürich (Kanton)* in the German version of the e-HDS is depicted. On the left, hyperlinks to the French (= *français*) and Italian (= *italiano*) versions of the article are shown and highlighted in blue. At the top right corner, users can click on hyperlinks to provide feedback (= *Rückmeldung*), open a *PDF copy* of the article, or *print* an article (= *drucken*). Left of the *Rückmeldung* hyperlink, the date *03/02/2015* is shown, which represents the date when the article was last updated by the *editorial office*. In the article text at the bottom of Figure 19, the term *Vorort* is highlighted in blue, which is a hyperlink to the article *Vorort* in the e-HDS.

Figure 19: First part of the article *Zürich (Kanton)* in the e-HDS (retrieved from Horisberger et al., 2015).

On the right in Figure 19, a field with a grey background is shown. In this field, a hyperlink to the publisher of the printed version of the HDS is highlighted in blue, and a text invites people interested in the illustrations and graphics of the article to order the printed version of the HDS.

Article title	DFI	Original language	Spatial, temporal, thematic information	Length
Zürich (Gemeinde)	X X X	-	ZH	51975
Zürich (Kanton)	X X X	-		192500
Zürich, Schlachten bei	X X X	d	1799	1925
Züricher, Gertrud	X X X	d	1871-1956	825
Züricher Post	X X X	d		825
Zürichgau	X X X	d		1925
Zürichsee	X X X	d	ZH, SZ, SG	7700

Figure 20: Selection of results for an e-HDS article title search with the query term *Zü* (retrieved from HDS, 2016b, adapted).

If a user chooses the *article search* (= *Artikelsuche*) option, the full article title (e.g., *Zürich*) or only the beginning of a title might be entered as a query (e.g., *Zü*), and all article titles beginning with the search term are displayed (e.g., *Zürich (Gemeinde)*, *Zürichgau*, *Zürichsee*)

in a list, including some metadata, as illustrated in Figure 20. By clicking on a title (e.g., *Zürich (Gemeinde)*), a new window containing the respective article pops up. By clicking on one of the crosses in the column labeled DFI in Figure 20, the article is opened in *German* (D), *French* (F), or *Italian* (I), respectively.

Metadata regarding the *original language* of the article, the *length* (i.e., number of characters), and *temporal*, *spatial*, or *thematic* information, are displayed, if available. As temporal metadata, the *birth date* and *date of death* for *biographies* (if known) as well as some dates for *thematic contributions* (e.g., date of *events*, *wars*) are listed. As spatial metadata, information about the *country* or the *canton* of many *geographical entities*, and some articles not belonging to the *geographical entities* category, is displayed. Multiple *cantons* or *countries* are mentioned for some articles. For example, for the article about the *municipality of Zurich* (= *Zürich (Gemeinde)),* the abbreviation for the *Canton of Zurich* (= ZH) is shown in Figure 20, whereas for the article about the *Lake Zurich* (= *Zürichsee*), the abbreviations for the cantons of *Zurich* (= ZH), *Schwyz* (= SZ), and *St Gall* (= SG) are listed because *Lake Zurich* is split between these three cantons. As thematic metadata, semantic tags for 76 of 3,067 total *thematic contributions* are available; for example, some newspapers have a tag called *newspaper* (= *Zeitung*).

However, many articles have no metadata (i.e., 5,917 articles in the *German version* of the e-HDS), and for the articles with metadata, either *spatial*, or *temporal*, or *thematic* information, but no combined information (e.g., spatial and temporal information), is displayed. Only 11.5% of the articles classified as *thematic contributions* are provided with metadata in the e-HDS, and therefore this is the category which is least specified with metadata.

Further spatial metadata is available in the data set which we obtained from the HDS, but which is not displayed in the e-HDS. This metadata class specifies a *spatial unit* or a *combination of spatial units* for all articles. However, this spatial information only specifies abbreviations of *countries* or *cantons* for the articles instead of numbers for statistical areas (e.g., official number of municipalities according to *Swiss Federal Statistical Office*) or exact geographic coordinates for places mentioned in the articles, for example.

Summary

The current version of the e-HDS provides only limited search and filtering options, while spatial, temporal, or thematic query possibilities are not yet implemented. The precision of spatial metadata is limited to the spatial scales of *cantons* and *countries*, and the temporal metadata are only complete for *biographies* and very little thematic metadata for the articles is provided. If metadata are available, only one dimension (i.e., spatial, or temporal, or thematic) is displayed to users of the e-HDS.

3.3 Future of the Historical Dictionary of Switzerland

The HDS has started a project called *New HDS* (= *Neues HLS*) which is the follow-up project and update of the e-HDS (HDS, 2016c). Suggestions for a new version of the e-HDS have already been drawn by Haber (2007, 2008). Haber (2008: 139) suggests the inclusion of all graphics, maps, and pictures of the printed HDS version in the new e-HDS and further argues that audio and video data should be incorporated and linked to the articles. In addition, the spatial, temporal, and thematic metadata of all articles should be systematically assessed in order to link these multimedia data to the e-HDS (Haber, 2008: 139-40). Haber (2008: 142-45) even argues for the inclusion of *social web* functionalities in order to potentially provide those interested in the HDS the opportunity to collaboratively elaborate new content for the e-HDS. However, there are inherent dangers in involving the general public in the writing process (e.g., *vandalism*, *advertisement*, *misuse*) and strategies to prevent quality loss (e.g., establish a *control instance*) must be developed (Haber, 2008: 143-45). Another idea of Haber (2008: 145-47) is to include HDS content on external websites and in web 2.0 applications (e.g., *mash ups*). For example, the *Google Maps API*[39] might be used to link HDS articles to locations on a map.

The HDS has presented the strategy of the *New HDS* in a management summary (HDS, 2016c), defining similar goals to those of Haber (2007, 2008), which can be briefly summarized as follows: the *New HDS* is a multimedia and multilingual online dictionary about the history of Switzerland aimed at the scientific community as well as the general public. The original contents of the HDS are adopted, thematically extended, and presented to interested users in new (interactive) search tools based on multidimensional (i.e., *spatial*, *temporal*, and *thematic*) metadata. Furthermore, the *New HDS* positions itself in an international and national information network, collaborating with other organizations in order to systematically interlink the *New HDS* with other data sources. The contents will still be produced by scientific experts only, and will be monitored and updated continuously in order to present findings according to the current state of research in Swiss history. In addition, text data, illustrations, audio and video documents, as well as information visualizations will be created and embedded on the HDS website, and optimized with regard to *usability* and *interoperability*. The launch of the *New HDS* is planned for 2017 (HDS, 2016c).

3.4 Data critique

In this section, the development and content of the HDS is critically analyzed. Most importantly, the HDS reflects the knowledge, findings, and interpretations of *authors*, *scientific advisors*, and people working at the *editorial office*; as such, it might not always represent the opinion of the majority of Swiss history experts. In addition, the selection of topics covered in the HDS as well as the length of articles is specified by the HDS. This implies that certain topics might be interesting to many people, but are not covered

[39] Link to Google Maps API: https://developers.google.com/maps/ (accessed June 2016)

or are only dedicated a few lines in the HDS, as decision-makers at the HDS rated them as being less relevant. Furthermore, the weights which have been assigned to specific time periods (e.g., 20% of the HDS content should be about the 20th century, as mentioned in *Section 3.1*), or to the spatial units in Switzerland, also influence the HDS content. The criteria to assign weights might be reasonable to the decision-makers of the HDS, while someone interested in time periods before the 17th century, for example, might not find as much information as expected because less recent time periods are covered in less detail.

Another point which influences the content of the HDS is the strict structural rules that had to be applied for writing the articles in order to keep them brief yet comprehensive. This is due to the fact that the HDS was initially planned to only be published as a series of printed books, and thus only limited space is available for each of the articles. As a consequence, there is often no space for sub-clauses explaining details regarding certain subjects covered in HDS articles (Morosoli, 2000: 15). Furthermore, for certain HDS articles, *authors*, *scientific advisors*, and the *editorial office* must choose examples to illustrate a subject as opposed to providing a complete overview of the subject, due to the restriction regarding article length. These choices influence the content of the HDS and could influence the results of our project. However, this point is not further evaluated in this thesis.

The temporal aspect is another factor influencing HDS content. The HDS project began in 1988, and the most recent articles for the printed book were written in 2014. Therefore, the HDS only represents the state of research within this specific time period. Furthermore, articles have been written and updated in different years within this time period, meaning that some articles might already be outdated while others represent the most recent findings in Swiss history research.

Human factors such as *knowledge* regarding a particular topic, *motivation*, and others, influence the HDS substantially as well. Thus, differences in the quality might occur among HDS articles due to the *authors* chosen to write articles, the *scientific advisers*, and the people responsible at the *editorial office*. Furthermore, the *translator* also influences the content of an article (Morosoli, 2000: 15). However, the fact that many people (i.e., *author*, *scientific adviser*, *editorial office*, *translator*) are involved in the writing and publishing process balances out these *human factors* to some degree.

Moreover, the *availability*, the *reliability*, and the *interpretation* of sources influence HDS content. In this context, *source criticism*, which analyzes who has created a source, the context, and the motivation of the creators, is relevant. By understanding the sources and creators, the *reliability* and *plausibility* of research results might be better assessed.

To summarize, the HDS represents one view on Swiss history (i.e., a *HDS view*) based on the state of research at the end of the 20th and the beginning of the 21st century. The selection of topics covered and people involved in the writing and publication process influence the contents of the HDS. In the context of this thesis, methods to analyze these factors (e.g., *source criticism*) are not applied, but limitations and potential

improvements for future work are discussed in *Chapter 7 – Discussion* and *Chapter 8 – Conclusions and Outlook*.

Summary

- The HDS is a multilingual dictionary about Swiss history and contains more than 110,000 articles about *geographical entities*, historically important *families*, *biographies*, and *thematic contributions*.
- The online version of the HDS offers limited search functionalities (i.e., *full text* and *article title* search) without an option to access articles by *spatial*, *temporal*, or *thematic* criteria.
- The *New HDS* will be launched in 2017 and will extend the current online version of the HDS through interactive search tools based on multidimensional metadata (i.e., *spatial*, *temporal*, and *thematic*) and multimedia content (e.g., audio and video documents, information visualizations).

As illustrated in this chapter, the most recent version of the e-HDS only provides *close reading* functionalities (i.e., reading HDS articles) and information seekers interested in the history of Switzerland can only currently access HDS articles by entering title or full text queries on the HDS website. No spatio-temporal or thematic search options are available, and *distant reading* is not supported, for example, by interactive visualizations which depict relationships or interconnections of spatio-temporal and thematic information in the HDS. We attempt to extend the existing *close reading* functionalities of the e-HDS, and, in the following chapter, illustrate our approach to provide coupled *distant* and *close reading* options to information seekers interested in Swiss history by visualizing spatialized displays of spatio-temporal and thematic information and relationships through interactive and exploratory web interfaces.

4 Methods

The overall workflow applied to this project is depicted in Figure 21. It is divided into three stages; (1) *geographic information retrieval* (GIR) to retrieve spatial, temporal, and thematic information from the *Historical Dictionary of Switzerland* (HDS) and store it in appropriate databases, (2) *spatialization* for computing spatio-temporal and thematic relationships and depicting them in spatialized displays, and (3) *geovisual analytics* (geoVA) as a method to create interactive and exploratory web interfaces which include spatialized displays, and which are presented to and empirically evaluated with target users (i.e., historians, people interested in *digital humanities*) of this project. The structure of this chapter follows these three stages.

All computational processes described in this chapter, as well as the second *think aloud study*, were run on a *Lenovo T430* laptop with an *Intel® Core™ i7-3530M CPU @ 2.90 GHz* , 8 GB installed memory (RAM), and *Windows® 7 64-bit* as the operating system. All databases were created, stored, accessed, and queried in *MySQL workbench*[40], a freely available visual user interface used to manage databases with the SQL query language[41].

4.1 Geographic information retrieval

We received all e-HDS articles in one XML formatted file[42] per language version (i.e., *German*, *French*, and *Italian*) and investigated the German version of the e-HDS. The reason for choosing the German version of the e-HDS is explained in *Section 3.2*. In order to retrieve information regarding space, time, and theme from the HDS articles, some preprocessing steps were necessary. All programming steps described in this section were run in *Python*[43] versions 2.7.5 and 3.3.2. *Python* was chosen as the programming language because it offers a standard library and several additional packages that are used in this project and are freely accessible online and widely used across the sciences. Since *Python* is used by a large community of programmers, online support for a wide range of topics is available.

[40] MySQL workbench: http://www.mysql.com/products/workbench/ (access June 2016)
[41] SQL: http://www.w3schools.com/sql/ (accessed September 2016)
[42] Introduction to XML: http://www.w3schools.com/xml/ (accessed June 2016)
[43] Python: https://www.python.org/ (accessed June 2016)

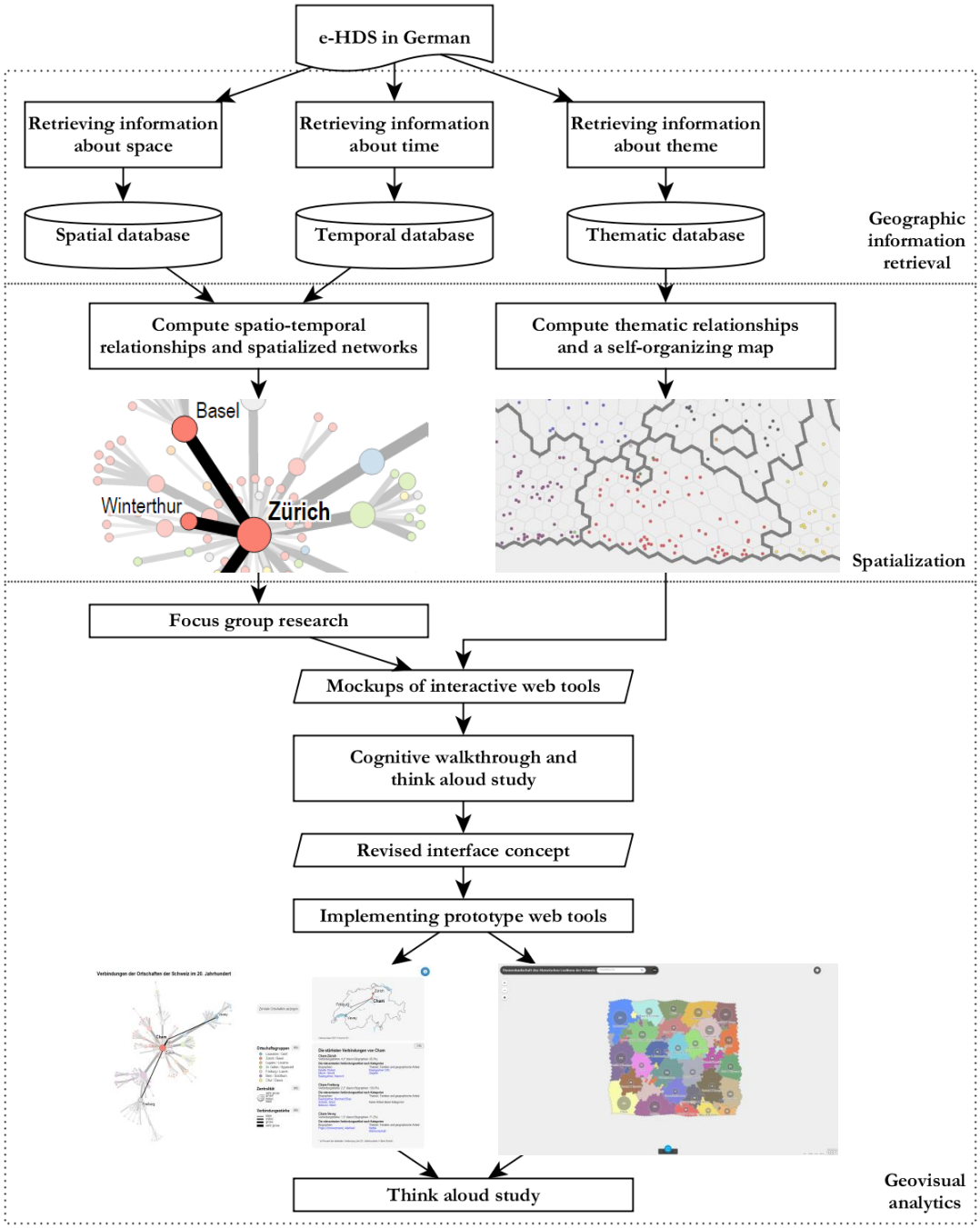
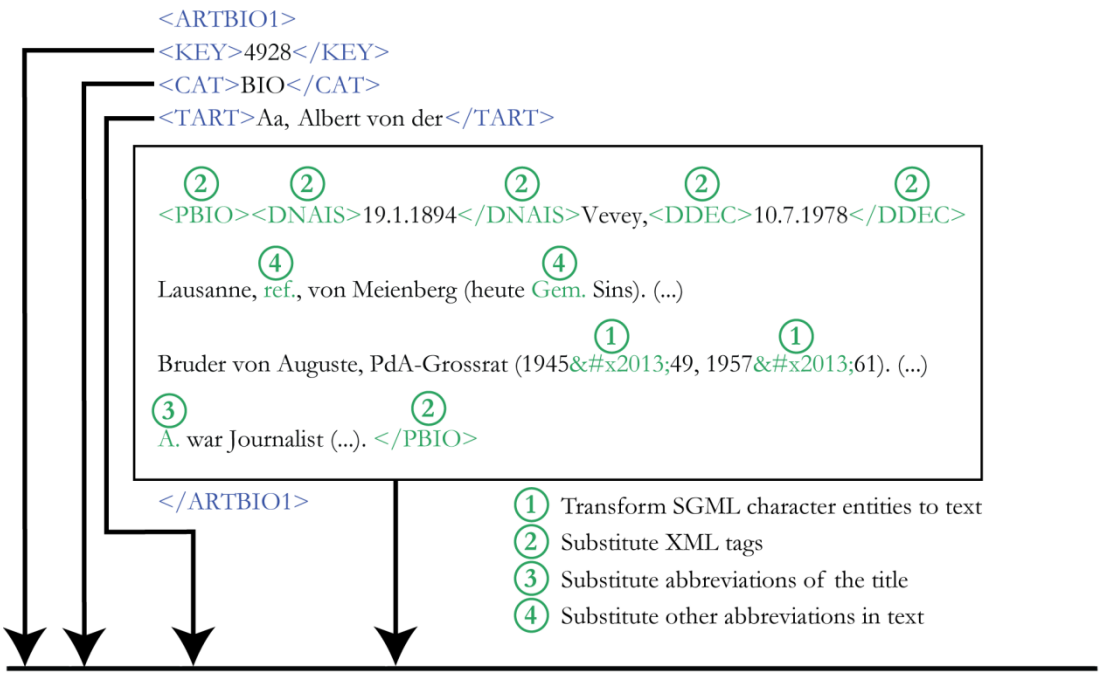


Figure 21: *Geographic information retrieval*, *spatialization*, and *geovisual analytics* depicted as part of the overall workflow.

In Figure 22, the preprocessing steps are illustrated with the example article *Aa, Albert von der* which is categorized as *biography* in the HDS. The original XML input file, and the output of the preprocessing steps (i.e., the table in Figure 22) are shown. For illustration purposes, only a part of the original article is depicted.



ID	Category	Title	Text
4928	BIO	Aa, Albert von der	geboren 19.1.1894 Vevey, gestorben 10.7.1978 Lausanne, reformiert, von Meienberg (heute Gemeinde Sins). (…) Bruder von Auguste, PdA-Grossrat (1945-49, 1957-61). (…) Aa, Albert von der war Journalist (…).

Figure 22: The preprocessing steps from XML input file (above) to database entry (below) illustrated with a part of the *Aa, Albert von der* article.

The XML input file is structured hierarchically and consists of start (i.e., <…>) and end (i.e., </…>) tags. Between the start and end tags in Figure 22, information is stored. The tag <ARTBIO1> indicates the beginning, </ARTBIO1> the end of the article. The number *4928* is shown between <KEY> and </KEY>, which is the *unique ID* of the HDS article. Between the CAT start and end tags, the *article category* is displayed, which is BIO (= *biographies*) for the example article in Figure 22. The TART tag contains the *article title*. All elements between <PBIO> and </PBIO> are part of the *article text*. In Figure 22 the *article text* is framed by a black rectangle.

The *article texts* (i.e., within the PBIO start and end tags) were preprocessed in order to be in a format which can be processed by the spatial, temporal, and thematic information retrieval algorithms applied in the next stages of the retrieval procedure (see *Subsections 4.1.1-4.1.3*). As the retrieval algorithms only accept raw text data without special characters as an input, *special characters* (i.e., number 1 in Figure 22) and *XML tags* (i.e., number 2 in Figure 22) had to be transformed. In addition, *abbreviations* (i.e., numbers 3 and 4 in Figure 22) had to be spelled out. This is important, as many abbreviations occur (e.g., for *article titles*, *adjectives*, and *adverbs*) in the HDS articles, and a lot of information would be missed without spelling them out. This is particularly relevant for the thematic information retrieval algorithm, which automatically assigns topics to the articles based on the words in the articles (see *Subsection 4.1.3*). Therefore, the more complete article texts are, the better the algorithm will perform.

We preprocessed the article texts in four steps: first, *SGML character entities*[44] had to be translated to *Unicode* characters, as illustrated in number 1 in Figure 22. In the e-HDS, *character entities* such as the *en dash* are encoded in *Hex Code* (e.g., *–*) and thus the *HTML parser library*[45] was employed to transform *character entities*. In Figure 22, the character entity *–* is thus converted to an *en dash* in the table in Figure 22.

In the second step, XML tags were substituted, which can be seen by number 2 in Figure 22. The DNAIS tag annotates the *birth date* and the DDEC the *date of death*. Therefore, we replaced the tags with the words *born* (= *geboren*) and *died* (= *gestorben*), respectively. Then, the PBIO start and end tags were removed. As a result of this preprocessing step, in column four of the table in Figure 22, the article text begins with *geboren* instead of <PBIO><DNAIS>, for example.

In the third step, we substituted the abbreviated titles in the article texts with the full title, as illustrated by number 3 in Figure 22. Article titles were abbreviated in the text to the initial letter of the title followed by a dot. Therefore, all patterns in the article text corresponding to the initial letter of the title plus a dot were replaced by the article title. As a result, in column four of the table in Figure 22, the fourth line begins with *Aa, Albert von der war Journalist* instead of *A. war Journalist* as in the original XML input file in Figure 22. Both Steps 2 and 3 in Figure 22 were performed by applying *regular expressions*[46] to identify search patterns (i.e., XML tags and abbreviated titles) in the article texts and substitute them accordingly.

In the final step, other abbreviations in the article texts were searched and substituted, as depicted by number 4 in Figure 22. For this purpose, the list of abbreviations published on the HDS website[47] was considered and transformed into a *dictionary*[48]. To this dictionary, further abbreviations were added manually because many *adjectives* and *adverbs* are abbreviated in the HDS, and thus the German word *reformiert* (= *reformed*), for example, is abbreviated to *ref.* in the article text in Figure 22. To locate such abbreviations, we searched the article texts for *characters plus dot* patterns (e.g., *ref.*) by applying *regular expressions* and manually added abbreviations to the dictionary. Abbreviations were considered if they occurred at least 350 times in the HDS (i.e., at least one occurrence in 100 HDS articles, on average). The full dictionary consisting of official (i.e., listed on HDS website) and manually added abbreviations was then used to replace the abbreviations in the articles (e.g., *ref.* = *reformiert*).

After these processing steps, the *unique ID* (i.e., KEY tag), the *article category* (i.e., CAT tag), the *article title* (i.e., TART tag), and the preprocessed *article text* of all 36,188 HDS articles were stored in a MySQL database in the format as illustrated in the table in

[44] SGML character entity: http://xml.coverpages.org/goldenti.html (accessed June 2016)
[45] HTML parser library in Python: https://docs.python.org/3.3/library/html.parser.html (accessed June 2016)
[46] Regular expressions in Python: https://docs.python.org/2/library/re.html (accessed June 2016)
[47] List of HDS abbreviations: http://www.hls-dhs-dss.ch/d/abkuerzungen/einleitung-zeichen (accessed June 2016)
[48] Python dictionary: https://docs.python.org/2/library/stdtypes.html#typesmapping (accessed June 2016)

Figure 22. The column *text* was used as an input for the information retrieval algorithms presented in the following subsections.

4.1.1 Retrieving spatial information

For the retrieval of spatial information, we needed to apply further preprocessing steps. First, we removed the *place of citizenship* (= *Bürgerort*) in *biographies* from the article texts. The *place of citizenship* is a status of Swiss citizens which is often neither the *place of birth* nor the *place of residence* of Swiss citizens. Instead, the *place of citizenship* is often obtained by birth from parents and often represents the *place of origin* of the family. Although the meaning was relevant in previous time periods due to the *rights and obligations* for a citizen at the *place of citizenship*, we judged it as unimportant to creating spatial relationships (as shown in *Subsection 4.2*) for people covered in the HDS because the HDS particularly contains *biographies* from relatively recent time periods. The *place of citizenship* is mentioned (if known) in the first sentence of *biographies*. For example, in the table in Figure 22, in the expression *Lausanne, reformiert, von Meienberg (heute Gemeinde Sins)*, the expression *Meienberg (heute Gemeinde Sins)* refers to the *place of citizenship*. This structure to name *places of citizenship* at the beginning of HDS articles is found in all *biographies*, and is therefore used to identify the *place of citizenship* by applying *regular expressions* and remove them from article texts. A short evaluation of 300 *biographies* showed that, in 99% of cases, the *place of citizenship* was identified and removed correctly.

A second preprocessing step considers the article category *geographical entities*. At the beginning of most articles about *municipalities* and *former municipalities* of Switzerland, the membership of the municipality to cantons and to districts (i.e., the administrative level between canton and municipality) is mentioned, as, for example, in the article *Dübendorf*, (illustrated in Table 2). The fourth column describes that *Dübendorf* is a *political municipality* of the *Canton of Zurich* (= *Politische Gemeinde Kanton Zürich*) and that *Dübendorf* is part of the *district of Uster* (= *Bezirk Uster*). However, we are not interested in analyzing information on the spatial hierarchies of cantons and districts in this project (see *Subsection 5.2.1*). Therefore, we identified and removed spatial expressions in the first sentences in the articles about municipalities and former municipalities by applying *regular expressions*.

Table 2: The *key*, *category*, *title*, and *text* of the article *Dübendorf*.

Key	Category	Title	Text
128	GEO	Dübendorf	Politische Gemeinde Kanton Zürich, Bezirk Uster. (…)

Following these additional preprocessing steps, an algorithm developed by Derungs and Purves (2014), described in detail by Derungs (2014), was run in a slightly adapted version. This algorithm fits the purpose of our project very well, as it is optimized for retrieving spatial information from unstructured German texts. Here, we will only discuss the parts of the algorithm that are relevant to this project. The code was

provided by courtesy of Curdin Derungs and consists of various scripts in the *Java* language[49].

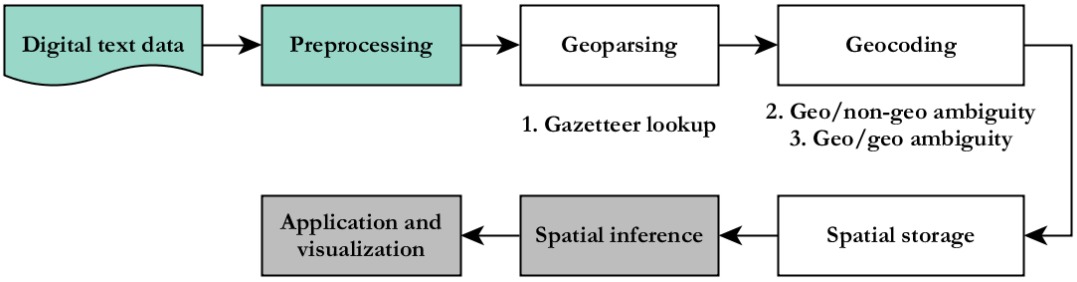


Figure 23: GIR model for retrieving spatial information from the HDS.

Figure 23 outlines the GIR schema for spatial information retrieval, which was introduced in *Subsection 2.1.1*. The first step from entire digital text documents to preprocessed text data has already been discussed, and therefore the respective boxes are colored in green. The *geoparsing* and the *geocoding* steps are explained in the following sections and enable the retrieval and disambiguation of spatial data. *Geoparsing* is based on a *gazetteer lookup* approach, and *geocoding* resolves *geo/non-geo* as well as *geo/geo ambiguities*. The retrieved spatial data is stored in a spatial database. The last two steps in Figure 23, which relate to *spatial inference* as well as *application and visualization*, were not applied, and therefore the respective boxes are colored in grey.

Geoparsing – Gazetteer lookup

The *geoparsing* and *geocoding* steps are described in pseudo code in Algorithm 1. Number 2 in Algorithm 1 represents the *geoparsing* step.

Algorithm 1: Pseudo code of the *geoparsing* and *geocoding* algorithm (adapted from Derungs, 2014: 77).

01:	**function geoparsing()**
02:	**Gazetteer lookup**: The text is parsed for terms that have similar wordings as toponyms, i.e., *potTop*. As a ground truth set of toponyms we use the SwissNames gazetteer. → *potTop*
03:	**Ambiguity**: All *potTop* are evaluated for geo/non-geo and geo/geo ambiguity. Geo/geo ambiguity is present if one *potTop* has several referent locations listed in SwissNames. Geo/non-geo ambiguity is existent if a *potTop* is tagged as a noun, and not a named entity, in the Tiger corpus. The result is a classification of all *potTop* into ambiguous (*ambTop*) and unambiguous toponyms (*unambTop*). → *ambTop/unambTop*
All of the following steps are only calculated for *ambTop* with geo/geo ambiguity. *ambTop* with geo/non-geo ambiguity are excluded at this stage. *UnambTop* are resolved as toponyms.	
04:	**Neighborhood**: For each *ambTop* with geo/geo ambiguity, we gather a set of neighboring *unambTop* (*neighTop*). Only *unambTop* within 100 words distance in text are considered. Each *neighTop* is associated with the word-count-distance from the respective *ambTop*. → *neighTop*
05:	**Euclidean Distance and Disambiguation**: We calculate a separate Euclidean distance for each referent location of an *ambTop* with geo/geo ambiguity and all *neighTop* (*D_ref*). The minimum *D_ref*, which is the referent location that is most proximate to the set of *neighTop*, is chosen to resolve the *ambTop* as a toponym.

[49] Java: https://www.java.com/ (accessed June 2016)

All HDS articles were processed word by word and compared to a gazetteer. We employed *SwissNames*[50] (version 2008) as a gazetteer, which consists of all toponyms found on topographic maps of Switzerland. For this project, *SwissNames* at a scale of 1:25,000 was chosen as this data set has the highest spatial resolution, containing 156,755 Swiss toponyms in total. If *SwissNames* toponyms were found in the articles, they were annotated as *potential toponyms*.

All toponyms in the *SwissNames* data set are georeferenced to a particular point location with a coordinate pair on a Swiss topographic map. Furthermore, metadata (e.g., *unique ID*, *feature type*.) for each toponym are also available. The *feature types* and their frequency of occurrence in the *SwissNames* database are shown in Table 3.

Table 3: *SwissNames* feature types.

Feature type	Total	Percent
Settlements (e.g., cities, municipalities)	62,323	39.8
Areas (e.g., forests)	62,160	39.6
Mountains	11,569	7.4
Rivers and lakes	7,550	4.8
Single objects (e.g., churches, castles)	5,096	3.2
Valleys	4,852	3.1
Passes	2,000	1.3
Roads and facilities	1,205	0.8
Total	156,755	100

Settlements and *areas* account for almost 80% of all toponyms in the *SwissNames* database, whereas *mountains*, *rivers and lakes*, *single objects*, *valleys*, *passes*, and *roads and facilities* account for the remaining 20% of toponyms.

Georeferencing all toponyms, including large areas (e.g., *cities*, *forests*) and linear features (e.g., *rivers*), to a particular point location instead of georeferencing them as areas or lines is one of the disadvantages of the *SwissNames* database (version 2008). Meanwhile, a new version of *SwissNames* has been published which includes all geometry types (i.e., points, lines, and areas) and might prove useful for future projects.

The list of *SwissNames* toponyms was manually extended. First, we added all labels for the Swiss cantons (e.g., *Kanton Zürich*) to the gazetteer as we realized after a first *geoparsing* run that cantons were identified as cities, very often by mistake, if the name of a canton is the same as the name of its capital city (e.g., *Kanton Zürich* vs. the city of *Zürich*). By separating the cantons from the cities, we aimed at improving the *precision* of the retrieval process for city toponyms. We evaluated this in a case study with *Zürich* and *Appenzell*. *Appenzell* is very strongly affected by this potentially wrong retrieval because there are two half-cantons (i.e., *Appenzell Innerrhoden* and *Appenzell Ausserrhoden*) potentially referring to the city of *Appenzell*. By adding the cantons to the gazetteer, we were able to improve the *precision* of the city of *Appenzell* from 0.53 to 0.70, and of the

[50] SwissNames on the website of the Swiss Federal Office of Topography swisstopo (version 2016): https://shop.swisstopo.admin.ch/en/products/landscape/names3D (accessed August 2016)

city of *Zürich* from 0.88 to 0.95 in a test collection, consisting of 100 randomly selected occurrences for each of the two toponyms in the HDS.

Another issue we faced was that many toponym names are stored in *SwissNames* only in one local language. For example, *Geneva* is stored as *Genève* (= French) and not as *Genf* (= German), because *Geneva* is located in the French-speaking part of Switzerland. However, in the *German* version of the HDS that we studied in this project, the German term *Genf* is used to describe *Geneva*. We thus extracted the titles of all *geographical entities* in *German*, *French*, and *Italian* and filtered article titles that are not identical in all three languages. We then checked whether *German* versions of the toponyms would be present in the *SwissNames* database. If not, the *German* versions of the toponyms were added to the gazetteer. Added toponyms which became relevant for further analyses are *Genf* (= *Geneva*), *Neuenburg* (=*Neuchâtel*), *Sitten* (= *Sion*), and *Pruntrut* (= *Porrentruy*).

Geocoding – Geo/non-geo ambiguity

The *potential toponyms* extracted in the *geoparsing* stage were then analyzed for *geo/non-geo ambiguity*, as illustrated by number 03 in Algorithm 1. Derungs (2014: 75) considered the TIGER corpus[51], which consists of approximately 50,000 sentences from the *Frankfurter Rundschau*, a German newspaper. From this corpus, Derungs (2014: 64-65) extracted 54,599 common German nouns. We used this TIGER nouns list and marked all *potential toponyms* in the HDS articles which are part of this list as *geo/non-geo ambiguous* toponyms, and thus did not investigate them any further as we expect these *geo/non-geo ambiguous* toponyms to primarily be used with their non-geographic meaning in the HDS (see the motivation for this approach in *Subsection 2.1.1*). We compared *SwissNames* to the TIGER corpus and determined that 3.5% of all unique toponyms in *SwissNames* are *geo/non-geo ambiguous*.

In addition, we checked for all major cities of Switzerland (i.e., capital cities of cantons and cities with population > 50,000), if they are part of the TIGER nouns list, and noticed that *Sitten* and *Zug* are on the list, because *Sitten* is a common German word for *morals* as well as the capital city of the *Canton of Valais*. *Zug* has several non-geographic meanings in German, and is also the capital city of the *Canton of Zug*. *Zug* used as a word for *train* is the most common non-geographic meaning of *Zug* in German. We analyzed 350 randomly selected occurrences of *Sitten* and *Zug* in article texts and evaluated the meaning of *Sitten* and *Zug* by reading the respective articles. We determined that the toponym *Sitten* was implied in 95.7% of the 350 occurrences of the term *Sitten* in HDS articles, while the non-geographic meaning of *Sitten* was implied in only 4.3% of occurrences. In contrast, in 66% of the 350 occurrences of *Zug*, the city of *Zug* was implied, while the non-geographic meaning of the word *Zug* was implied in 34% of occurrences. As a consequence, we left *Zug* on the TIGER nouns list and thus marked *Zug* as *geo/non-geo ambiguous*, while *Sitten* was removed from the TIGER nouns list (i.e., *Sitten* is not considered *geo/non-geo ambiguous*).

[51] TIGER corpus: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html (accessed June 2016)

Geocoding – Geo/geo ambiguity

Potential toponyms were also analyzed for *geo/geo ambiguity*, as illustrated by number 03 in Algorithm 1. *Geo/geo ambiguity* is given if there are several *referent locations* in *SwissNames*, which implies that more than one toponym have the same name but refer to different locations (e.g., *Rüti* refers either to *Rüti* in the *Canton of Zurich* or *Rüti* in the *Canton of Glarus*). This holds true for 43% of all toponyms in *SwissNames* (Derungs, 2014: 61). For all *geo/geo ambiguous* toponyms the procedure illustrated by numbers 04 and 05 in Algorithm 1 was applied. First, for *geo/geo ambiguous* toponyms, all neighboring toponyms (*neighTop*) which are neither *geo/non-geo* nor *geo/geo ambiguous* and which are within a maximum distance of +/- 100 words in the article text, were considered. In Derungs (2014), 200 words is applied as a threshold for distance, but for longer text documents compared to the HDS articles. Therefore, we reduced the threshold to 100 words. For each of the *referent locations* of a *geo/geo ambiguous* toponym (*ambTop*), the Euclidean distance to all unambiguous *neighTop* was calculated. The Euclidean distance to each *neighTop* was weighted with the *inverse text distance* (i.e., *1 divided by text distance*). Therefore, an unambiguous *neighTop* which is closer in the article text to the *geo/geo ambiguous* toponym has a stronger influence on disambiguating the *geo/geo ambiguous* toponym. Derungs (2014: 75) motivates this weighting with the *first law of geography* (Tobler, 1970) and expects that this law holds true for the use of toponyms in texts as well. Therefore, Derungs (2014: 75) assumes that the shorter the text distance between two toponyms (e.g., two toponyms co-occur in the same sentence), the higher the probability that they are located close together in geographic space.

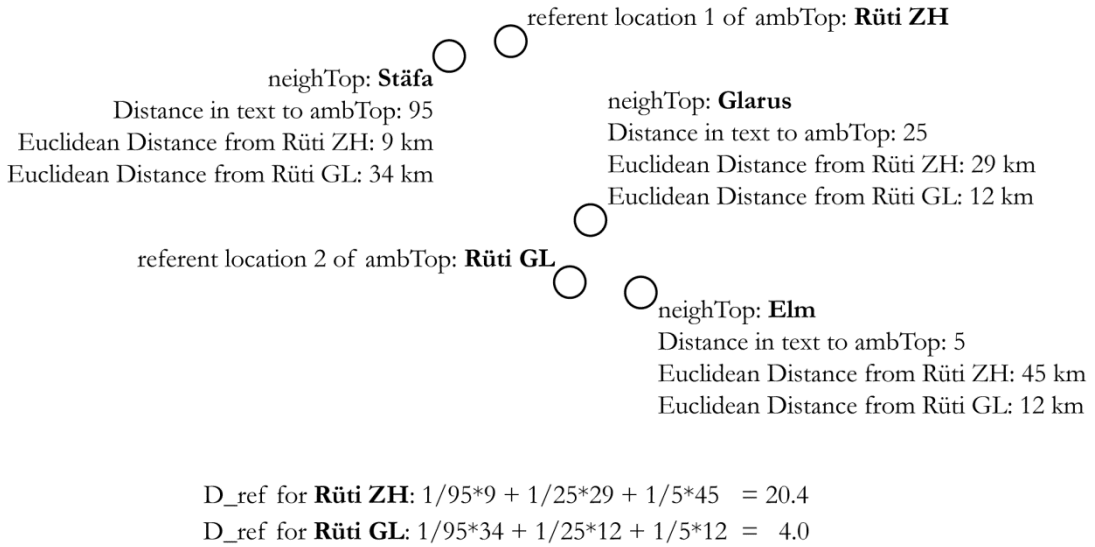


Figure 24: Resolving *geo/geo ambiguity* of *Rüti*.

Figure 24 illustrates an example of resolving *geo/geo ambiguity*. We assume that two *referent locations* exist for *Rüti*: *Rüti* in the *Canton of Zurich* (ZH) and *Rüti* in the *Canton of Glarus* (GL). Furthermore, we assume that *Stäfa*, *Glarus*, and *Elm* are within the defined maximum distance of 100 words to *Rüti* in the example article text. As *Glarus* and *Elm* are closer in Euclidean distance to the referent location *Rüti GL* than to *Rüti ZH*, the *total distance* (*D_ref*) to *Rüti GL* is lower than to *Rüti ZH* for these *neighTops*. Although

Stäfa is closer to *Rüti ZH* than to *Rüti GL* in Euclidean distance, the influence of *Stäfa* to both *D_ref* values is very low, as *Stäfa* is much further away in the article text to *Rüti* (i.e., 95 words) than *Glarus* and *Elm* (i.e., 25 and 5 words, respectively). Therefore, the *geo/geo ambiguous* toponym *Rüti* is resolved to *Rüti* in the *Canton of Glarus* in Figure 24.

Spatial storage

All unambiguous and disambiguated toponyms, their *unique IDs*, their *geographic coordinates*, and the *IDs* of the articles in which they occur were stored in a MySQL database table.

Having illustrated the retrieval of spatial information from the HDS, we now turn to the retrieval of the second geographic information dimension: time.

4.1.2 Retrieving temporal information

We retrieved temporal information from the HDS articles by employing *HeidelTime* 2.0.1[52]. *HeidelTime* fits well for this project, as it is an open-source multilingual tool to automatically retrieve temporal information from unstructured and semi-structured texts, with support for German texts (Strötgen and Gertz, 2013, Strötgen, 2015). In addition, it is optimized for retrieving historic dates and is applicable to different document types such as *narrative-style* (e.g., *Wikipedia*, *HDS*) or *news-style* texts (Strötgen et al., 2014).

All HDS article texts were used as an input to *HeidelTime*. The first step in the *HeidelTime* pipeline is the application of a *part-of-speech* tagger (POS). *HeidelTime* incorporates *TreeTagger*[53] as a POS tagger by default. For our project, *TreeTagger 3.2* was employed (Schmid, 1994, 1995). As a result of this preprocessing step, all HDS articles were *tokenized* (i.e., split into words, symbols, and numbers) and annotated with POS (e.g., noun, verb, numeral) and *lemma* information. *Lemma* is the *base form* of a word; for example, *package* is the *lemma* for *packages* and *receive* is the *lemma* for *receives*. Table 4 provides an example of a *TreeTagger* output for the sentence "We travel to France" in German (= "Wir reisten nach Frankreich"). The POS tags imply *personal pronoun* (PPER), *finite verb* (VVFIN), *preposition* (APPR), *named entity* (NE), and *punctuation mark* ($).

Table 4: POS tagging and lemmatization.

Token	POS tag	Lemma
Wir	PPER	wir
reisten	VVFIN	reisen
nach	APPR	nach
Frankreich	NE	Frankreich
.	$.	.

[52] GitHub repository of HeidelTime: https://github.com/HeidelTime/heideltime (accessed June 2016)
[53] TreeTagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (accessed June 2016)

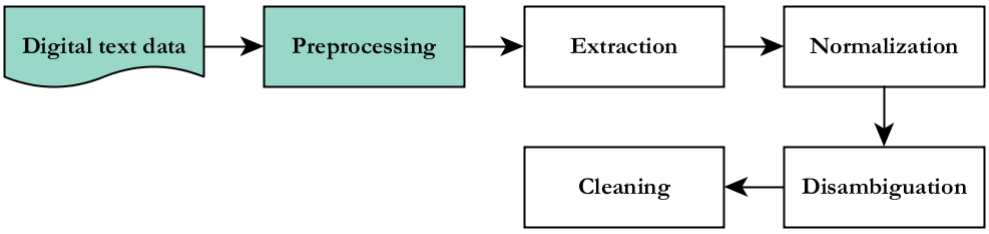


Figure 25: GIR model for retrieving temporal information from the HDS.

After this preprocessing step, *HeidelTime* applies *extraction*, *normalization*, *disambiguation*, and *cleaning* methods, as introduced in *Subsection 2.1.2* and visualized in Figure 25. The first step from digital text documents to preprocessed data has already been described; therefore, the respective boxes are colored in green.

- **Extraction**. The extraction in *HeidelTime* follows a *rule based* approach and applies primarily *regular expressions* to identify temporal information in the HDS articles (Strötgen and Gertz, 2013: 280). As an output of the extraction process, all *temporal references* in the HDS are annotated.

- **Normalization**. Extracted temporal references are normalized in *HeidelTime* following the *TimeML* standard using *Timex3* tags, which implies that all references are classified as either *date* (e.g., *September 27, 2016*), *time* (e.g., *5 p.m.*), *duration* (e.g., he worked for the company *for five years*), or *set* (e.g., *weekly*) in a standardized format as for example *2016-09-27* for *September 27, 2016* (Pustejovsky et al., 2005). The normalization in *HeidelTime* follows a *rule based* approach (Strötgen and Gertz, 2013: 282).

- **Disambiguation**. *Ambiguous* and *underspecified* temporal references are resolved in the *disambiguation* stage (Strötgen and Gertz, 2013: 287-88). For example, if *5 p.m.* is mentioned in an article, this reference is normalized to *XXXX-XX-XXT17:00* and thus *year*, *month*, and *day* are missing. Therefore, *HeidelTime* checks the availability of temporal references before or after in the text and attempts to specify the unspecified reference with this information.

- **Cleaning**. In the *cleaning* stage, temporal references which were extracted first, but identified as not being temporal references in the *normalization* stage, are removed (Strötgen and Gertz, 2013: 288). To identify *non-temporal* references, the result of the POS tagging is used. For example, the POS tagger identifies that in the expression *1990 miles*, a four-digit number is followed by a plural noun. Then, *HeidelTime* annotates this as a *non-temporal* reference in the *normalization* stage, as a number followed by a plural noun is a common pattern for a number used in a *non-temporal* context. In the *cleaning stage*, such references are removed.

For this project, only temporal references which were classified as *date* or *time* (if the year is specified) were considered for further analyses, as we were interested in grouping temporal references to *centuries* and only references from these two classes can be classified into *centuries*. The allocation of temporal references to *centuries* was completed

manually by sorting the normalized values of the references in *Microsoft Excel* by year and assigning the respective century. The reason for grouping the temporal references to *centuries* is detailed in *Subsection 4.2.1*.

All *normalized values of temporal references*, the assigned *centuries*, and the *IDs* of the articles in which they occur were stored in a MySQL database table.

Up to this point we have illustrated the retrieval of spatial and temporal information. In the following subsection, we consider the retrieval of the third dimension of geographic information: theme.

4.1.3 Retrieving thematic information

In order to retrieve thematic information, we decided to apply *probabilistic topic modeling* to the HDS articles. We chose *probabilistic topic modeling* as it is a method to automatically create topics from large unstructured and semi-structured text archives, and to automatically assign topic weights to the documents in the text archive (as introduced in *Subsection 2.1.3*). Therefore, *probabilistic topic modeling* supports understanding and learning about the thematic structure of a text archive by revealing the latent structure behind a collection of text documents. This is particularly helpful if the thematic structure of a text archive is not known, as it holds true for our project as we have illustrated in *Section 3.2* that only a small amount of thematic metadata is available in the HDS. Furthermore, *probabilistic topic modeling* supports the automatic generation of a bimodal matrix (i.e., *article-topic matrix*). A bimodal matrix is needed as an input to create *self-organizing maps*, as detailed in *Subsection 4.2.2*.

We applied *probabilistic topic modeling* to the 3,067 *thematic contributions* articles of the HDS. Choosing the *thematic contributions* articles is based on several factors: we aim to depict the *thematic landscape* of important themes in Swiss history (see *Subsection 4.2.2*), as target users in a *focus group meeting* (see *Subsection 5.3.1*) required access to Swiss history from a thematic point of view in addition to spatio-temporal access possibilities. The *thematic contributions* articles cover important themes surrounding Swiss history and contain a wealth of thematic information (see *Subsection 5.1.3*), thus were determined to be suitable for creating the *thematic landscape*. The other article categories (i.e., *biographies*, *families*, and *geographical entities*) focus on spatial and temporal information (see *Section 5.1*), and were thus considered less relevant to creating the *thematic landscape* of Swiss history.

We applied the *Machine Learning for Language Toolkit* 2.0.8 (MALLET)[54] to calculate the *probabilistic topic modeling* of the 3,067 *thematic contributions* articles in the HDS. MALLET is a *Java*-based open-source software and is implemented with a LDA technique and a fast *Gibbs sampling* implementation (McCallum, 2002). Readers interested in details regarding the algorithm are referred to Steyvers and Griffiths (2007). We decided to employ MALLET, as it is a common *topic modeling* (TM) tool and because MALLET provides quantitative model fit measures in the TM output, which allow for evaluation and comparison between different TM results (McCallum, 2002). We also illustrate how

[54] MALLET: http://mallet.cs.umass.edu/index.php (accessed June 2016)

quantitative model fit measures were utilized in the evaluation of this thesis (see *Section 6.3*).

Before we entered the *thematic contributions* articles in MALLET, we preprocessed them by applying *TreeTagger 3.2*, which was introduced in *Subsection 4.1.2*. As an output of *TreeTagger*, the article texts were *lemmatized*. *Lemmatization* is useful, as words have different grammatical forms due to *tenses* or *numbers* (i.e., *singular* or *plural*), for example. Using the *lemmas* of words as opposed to *non-lemmatized* and *inflectional forms* of words for TM has the advantage of a word occurring in different *inflectional forms* being treated as one word in the *topic* creation process. Another possibility to treat *inflectional forms* would be to apply *stemming*. However, *stemming* typically truncates the ends of words; therefore, words with different meanings but the same *stem* are no longer differentiable after removing *derivational affixes* (Manning et al., 2009e). For example, *stemming* would reduce all of the following words to the stem *oper*: *operate*, *operating*, *operates*, *operation*, *operative*, *operatives*, and *operational* (Manning et al., 2009e). This would result in a loss of detail and accuracy for TM (Manning et al., 2009e, Keim et al., 2013: 53). Therefore, we decided not to use *stemmed*, but rather *lemmatized* article texts.

We had to apply further preprocessing methods due to the influence of spatial and temporal information. Although we only considered *thematic contributions* articles with few spatial and temporal information compared to the other article categories (see *Section 5.1*), we realized in the preliminary runs of TM that spatial and temporal information does indeed influence the TM solution significantly. Some *topics* were created primarily based on specific time periods or geographical areas. However, as we were interested in identifying the thematic content of the articles and wished to retrieve spatial, temporal, and thematic information from the HDS articles separately, we decided to remove all spatial and temporal information from the articles before running MALLET. For this reason, we searched any *digits* occurring in the articles by applying *regular expressions* and removed them accordingly. Following this procedure, we deleted all *dates*, *times*, and other *numbers* from the article texts as we aimed at excluding the influence of numerical information for *topic* creation. In the next step, we created a *dictionary* consisting of spatial and temporal *stop words*. We considered all article titles of the *geographical entities* article category as spatial *stop words*. Furthermore, we added the names of all *cantons* as well as the toponym *Switzerland* (= *Schweiz*) to the spatial *stop word* list. Additionally, we considered *months of the year* as well as other common temporal expressions such as *century* (= *Jahrhundert*) and *BC (= vor Christus)* as temporal *stop words*. In addition, we included the *Natural Language Toolkit* (NLTK)[55] stop word list which contains 231 high-frequency words in German that have little lexical content and fail to distinguish texts from one another such as *the* (= *der*, *die*, *das*), *be* (= *sein*), or *to* (= *zu*) (Manning et al., 2009e). The dictionary of spatial, temporal and NLTK stop words was then used to automatically remove the stop words from the HDS articles by applying a *Python* script.

[55] NLTK stop word list: http://www.nltk.org/book/ch02.html (accessed June 2016)

After these preprocessing steps, we imported the 3,067 *thematic contribution* articles into MALLET. In MALLET, we applied default parameters and decided on the *hyperparameter optimization* method. *Hyperparameter optimization* allows some *topics* to be more prominent than others in the TM output (Wallach et al., 2009a). Furthermore, we had to define the number of *topics* to be created by MALLET. We decided to follow a combined quantitative and qualitative approach to find an appropriate number of *topics*, as suggested by Chang et al. (2009) and Mimno et al. (2011). In order to quantitatively assess an appropriate number of *topics*, we considered the *log likelihood/token* (LLT), which is shown in the MALLET output. The LLT indicates the model fit and thus how accurate the choice of a number of *topics* is (Griffiths and Steyvers, 2004, Wallach et al., 2009b: 1106-07). The higher the LLT value, the better the TM solution. By comparing the LLT of different TMs, one may choose the solution with the highest LLT value. In our project, we did TM runs for different numbers of *topics* and identified two solutions with good LLT values: 24 *topics* and 30 *topics*. In order to qualitatively assess different TM solutions, we visually compared them to one another and decided for the 30 *topics* solution. The evaluation procedure, including the quantitative and visual approaches, is covered in detail in *Chapter 6 – Evaluation*.

The output of the TM process is an *article-topic matrix* which assigns a vector to each article. The vectors describe the probability distribution of topics over articles (see Figure 5). This thematic data was stored in a MySQL database.

In this section, we have illustrated the retrieval of spatial, temporal, and thematic information from the HDS articles and thus our approach to address *Research Question 1* of this thesis (see *Section 1.3*). In the following section, we present how we addressed *Research Question 2*.

4.2 Spatialization

The spatial, temporal, and thematic data we retrieved from the HDS articles were transformed, reorganized, and visualized in spatialized displays in the next step. The *spatialization framework* and *spatialized displays* were introduced in *Section 2.2*. In this section, we focus on how we depict the spatio-temporal and thematic information in the HDS articles in *network visualizations* and *self-organizing maps* which were found to be adequate spatialized displays in the context of this research project (see *Section 2.2*). We thus provide an answer to *Research Question 2* (i.e., *how can we spatialize uncovered spatio-temporal and thematic structures and interconnections extracted from unstructured or semi-structured text archives in the humanities*), and present a spatio-temporal and thematic *distant reading* approach, which is applied to our case study (i.e., HDS). The approach presented in this section is inspired by Salvini (2012) and Salvini and Fabrikant (2016).

In the following subsections, we first detail the approach for computing spatio-temporal relationships and depict them in *network visualizations* in *Subsection 4.2.1*. Then, the computation of thematic relationships and visualization in *self-organizing maps* is detailed in *Subsection 4.2.2*.

4.2.1 Relationships between toponyms over time

In *Section 2.2*, we illustrated that using network visualizations is an effective approach to highlight relationships and interconnections between data items. We further introduced Salvini and Fabrikant's (2016) work on how to visualize a network spatialization of world cities based on the *Wikipedia* hyperlink structure. We decided to follow Salvini and Fabrikant's (2016) approach to compute and display spatial relationships in network visualizations, but added the temporal information dimension to networks in order to allow information seekers in the humanities to analyze spatial structures and relationships in and over different periods of time. As we selected a text archive about history for this thesis, incorporating temporal information is particularly important, since time is a very important factor in history (e.g., what happened and who lived when). The incorporation of network visualizations in web interfaces is detailed in *Subsection 4.3.2*.

We followed a two-step approach to compute and display relationships between toponyms over time according to the *spatialization framework* outlined by Fabrikant and Skupin (2005): first, spatio-temporal relationships are computed based on the spatial and temporal information retrieved from the HDS. Second, the spatio-temporal relationships are depicted as links between nodes in network visualizations. The entire procedure applied in this thesis is illustrated in Figure 26. With the exception of the final step in Figure 26, all steps consider the transformation of the spatial and temporal data retrieved from the HDS and the computation of the spatio-temporal networks. The final step represents the visualization of the *spatialized networks*. This subsection is structured following the steps in Figure 26; therefore, we first explain the selection of toponyms.

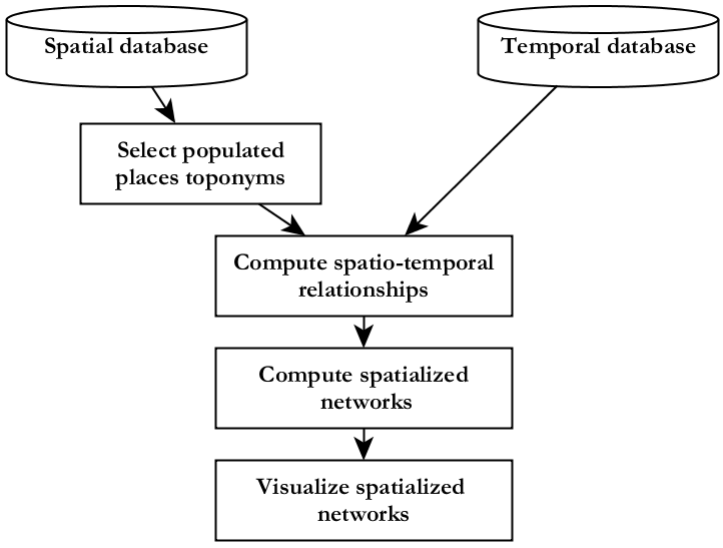


Figure 26: *Spatialized network* approach.

Selection of populated places toponyms

From the spatial database, we selected only toponyms of the feature type *settlements* in *SwissNames* (i.e., *villages*, *municipalities*, and *cities*). The reason for choosing *settlements* was that we are particularly interested in spatial links between *populated places*, motivated by Salvini and Fabrikant (2016), and thus *feature type classes* such as *mountains*, *rivers* and *lakes*, and *passes* were not considered. Contrary to Salvini and Fabrikant (2016), who established a *world city network*, we analyzed *populated places* on a *national* instead of a *global* level.

Numbers and statistics from the retrieved data (e.g., toponyms per *feature type* class) are presented in *Subsection 5.1.1*.

Computing spatio-temporal relationships

In the next step, we incorporated spatial as well as temporal information in order to compute spatio-temporal relationships. Spatial relationships were assessed by analyzing co-occurrences of toponyms in HDS articles. This approach is inspired by Hecht and Raubal (2008) and Salvini and Fabrikant (2016), who applied a similar idea to *Wikipedia* data. In their approach, they investigated the hyperlink structure of *Wikipedia* to define spatial and semantic relationships (see *Subsection 2.2.3*). Salvini and Fabrikant (2016: 233-34) defined a relationship between two cities if both are hyperlinked in the same *Wikipedia* article. The more the two cities are hyperlinked in the same articles, the stronger their relationship. We adopted this idea and further incorporated temporal information to specify spatio-temporal relationships.

As mentioned in *Subsection 4.1.2*, we assigned all temporal references in the HDS to *centuries*. This choice was based on discussions with historians involved in this project (e.g., target group members which participated in empirical studies). The two main arguments for selecting *centuries* emerged in these discussions: first, for previous time periods, little data is available in the HDS (i.e., before the 19th century), as illustrated in *Subsection 5.1.2*. Therefore, selecting a higher temporal resolution than *centuries*, such as *decades*, would cause only few spatio-temporal relationships to be depicted in *spatialized networks* about *decades*, and thus, an incomplete network of Swiss toponyms would be drawn. Second, aggregating to historical meaningful time periods (e.g., *Ancien Régime* in Swiss history which lasted from 1712 to 1798) would also cause incomplete networks, as these time periods often do not correspond with full centuries and therefore the same problem would occur as for aggregating to *decades*. Due to these reasons, we decided to aggregate the HDS data at the *centuries* level.

The computation of spatio-temporal relationships in the HDS is illustrated in Figure 27. We decided to calculate the relationships based on a spatial and a temporal weight. The spatial weight expresses how relevant toponyms are for toponym relationships in HDS articles and the temporal weight is used to assign the spatial relationships to *centuries*. The *Okapi BM25* formula is detailed in Manning et al. (2009c) and is a common algorithm for *probabilistic information retrieval*. It assigns a relevance score to terms in documents. This score is dependent on several factors: the *frequency of a term* in a document, the

length of a document compared to the *average document length* in a corpus, the *number of documents* in which the term occurs, and two *free parameters* (i.e., *k1* and *b* in Figure 27) (Manning et al., 2009c). The largest difference to other common methods applied in IR, such as the *tf-idf* algorithm (see *Subsection 2.1.3*), is the incorporation of the *article length*, which we deemed relevant for this project because article lengths in the HDS vary significantly, as demonstrated in Table 1. The *Okapi BM25 formula* in Figure 27 states that a toponym is most relevant if (a) it occurs very often in an article, (b) the *article length* is below average, and (c) if the toponym is rare in the entire corpus (i.e., the toponym occurs in very few articles).

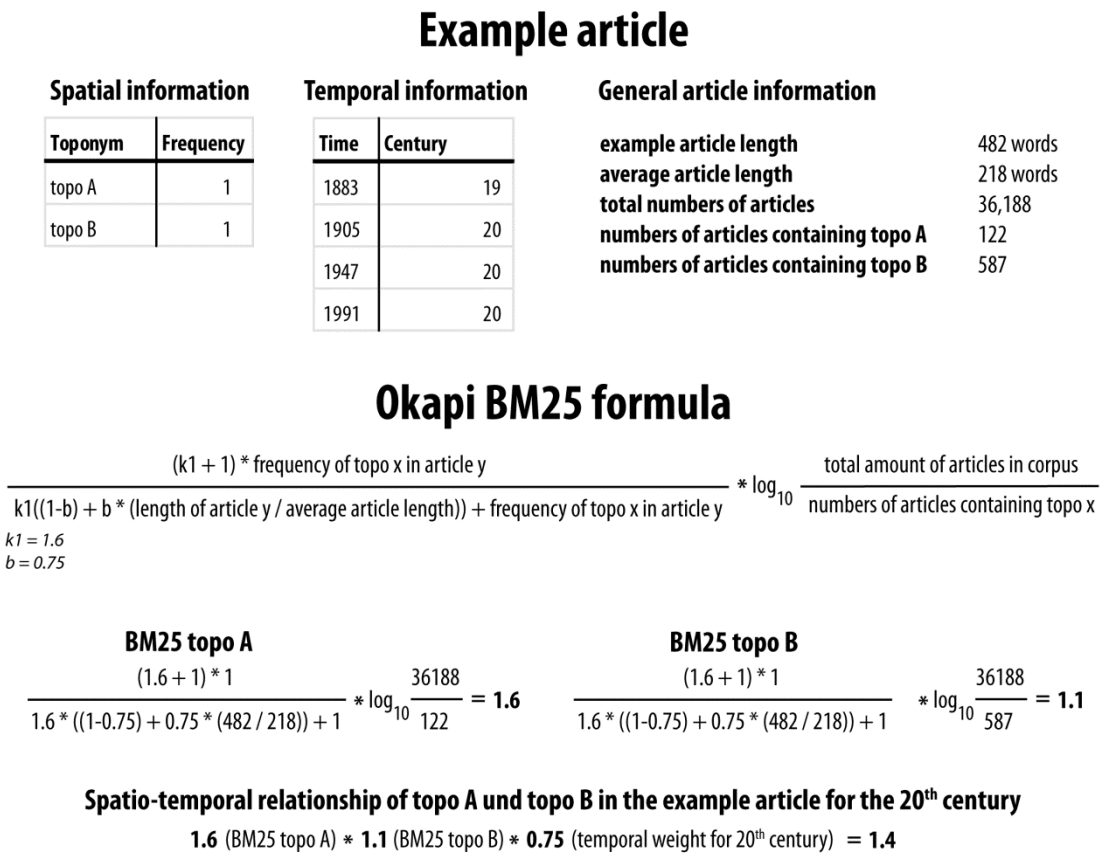


Figure 27: *Okapi BM25* illustrated with an example article.

In Figure 27, an example article is shown to illustrate the *Okapi BM25*. The two tables in the *example article* section in Figure 27 show the spatial (left) and the temporal information (middle) in the article. They illustrate that both toponyms (i.e., *topo A* and *topo B*) occur once in the example article, and that 75% of the temporal references in the article are related to the 20th century while 25% are related to the 19th century. In the *Okapi BM25 formula* section, the formula of the *Okapi BM25* is shown; below, the calculation of the *Okapi BM25* score for both toponyms is illustrated (i.e., *BM25 topo A*, *BM25 topo B*). The *Okapi BM25* score of *topo A* (i.e., 1.6) is higher than the *Okapi BM25* score of *topo B* (i.e., 1.1). This is due to the rareness of *topo A* in the entire HDS corpus when compared to *topo B*, as *topo A* only occurs in 122 HDS articles, compared to 587 articles for *topo B*. All other variables of the *Okapi BM25* are equal for *topo A* and

topo B (i.e., *frequency*, *article length*). At the bottom of Figure 27, the combined spatio-temporal score for the toponym relationship is shown. First, both *Okapi BM25* scores are multiplied and thus represent the spatial weight. Then, the spatial weight and the temporal weight are multiplied. We chose the 20th century to analyze in Figure 27. The temporal weight for the 20th century is 0.75 because 75% of the temporal references in the example article are about the 20th century. The calculation illustrated in Figure 27 was applied to all toponym relationships in all HDS articles. Then, the spatio-temporal relationships were summed up for each pair of toponyms and separated by *centuries*. All calculations were completed by conducting queries in *MySQL Workbench*.

The free parameters *k1* and *b* in Figure 27 were set to 1.6 and 0.75, respectively, based on the recommendations of Manning et al. (2009c). Furthermore, we decided to consider relationships of toponyms for a specific century only if the *temporal weight* of an article for this *century* is at least 50%. Thus, we assume that approximately half of the article's content or more is about the respective *century*. In addition, we only considered the strongest 90% of spatio-temporal relationships per *century*, and only relationships which are based on at least three articles in which two toponyms co-occur. Applying these filtering criteria excluded very low scoring spatio-temporal relationships. The influence of applying these filtering criteria to the spatio-temporal networks is evaluated in *Section 6.2*.

This developed method differs significantly from the method proposed in Salvini and Fabrikant (2016), as it incorporates not only spatial information, but also temporal information. The spatial weighting in Salvini and Fabrikant (2016) is similar to our approach (e.g., including *article length*), but was computed slightly differently as it was optimized for *Wikipedia* articles. Interested readers are referred to Salvini (2012) and Salvini and Fabrikant (2016).

Computing spatialized networks

The spatio-temporal relationships between toponyms (i.e., one network for each *century*), were then transferred as *text files* from *MySQL Workbench* to *txt2Pajek*[56]. The *txt2Pajek* tool transforms *text files* into *.net files*, which are readable by network analysis software such as *Pajek*[57] or *Network Workbench* (NWB)[58] (Pfeffer et al., 2013). In the next step, we employed *NWB 1.0.0* which is a toolkit for processing, analyzing, modeling, and visualizing large network data sets, and thus provides all functionalities needed to compute and depict the *spatialized networks* (NWB Team, 2006). First, we computed the *strength* of each toponym in NWB, which is the *sum of spatio-temporal relationship weights* of a toponym to all other toponyms in the network and thus indicates the *centrality* of a toponym (Wasserman and Faust, 1994: 178-83). Then, we computed the *community detection algorithm* presented by Blondel et al. (2008). This algorithm delineates toponym clusters. A toponym cluster consists of densely-connected toponyms within a cluster and weak connections to toponyms outside a cluster. The advantage of

[56] txt2Pajek: http://www.pfeffer.at/txt2pajek/ (accessed July 2016)
[57] Pajek: http://mrvar.fdv.uni-lj.si/pajek/ (accessed July 2016)
[58] Network Workbench: http://nwb.cns.iu.edu/ (accessed July 2016)

delineating clusters with this algorithm compared to other common grouping and clustering methods (e.g., *k-means* algorithm, see *Subsection 2.2.3*) is that the optimal number of clusters is automatically determined by the algorithm and thus does not need to be defined in advance. Furthermore, we applied the *pathfinder network scaling* algorithm in NWB (Dearholt and Schvaneveldt, 1990). *Pathfinder network scaling* is a very common method in *information visualization* for the reduction of complex and large networks to the structurally most relevant relationships (Börner et al., 2003: 201-03).

Visualizing spatialized networks

In the next step, we depicted the spatio-temporal relationships in network visualizations according to the *spatialization framework* and followed the recommendations by Salvini and Fabrikant (2016) regarding the *generalization* process. First, *semantic generalization* was performed: the toponyms were interpreted as *loci* and relationships between them as *trajectories*. Second, *geometric generalization* was applied: the geometric primitive *point* (i.e., *node*) was assigned to the semantic primitive *locus*, and *line* (i.e., *edge*) to *trajectory*. Furthermore, the visual variables *size*, *color hue*, and *color value* were applied to depict the *strength*, the *Blondel community*, and the *strength* of the toponym relationship (i.e., *weight*), respectively.

Toponym characteristics

Node	Community	Strength
topo A	blue	high
topo B	blue	medium
topo C	red	high
topo D	blue	medium

Toponym relationship

Node 1	Node 2	Weight
topo A	topo B	medium
topo A	topo C	strong
topo A	topo D	medium

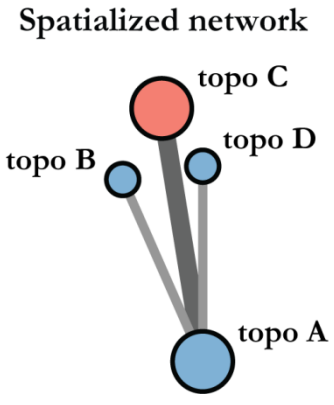


Figure 28: Generalized *spatialized network* visualization.

In Figure 28, an example of a *spatialized network* visualization, according to the aforementioned procedure is depicted. The *toponym characteristics* are represented by their *color hue* and the *size* of the respective nodes in the network visualization. *Topo C* is the only node which was assigned to the community *red* and thus is colored differently.

Topo A and *topo C* are more central (i.e., have a higher *strength* value) than *topo B* and *topo D* and thus are larger. The *toponym relationship* values (i.e., *weight*) are represented by *color value* and *edge size*. The edge between *topo C* and *topo A* is larger and represented in a darker color value than all other edges, as this relationship is the strongest in Figure 28.

Inspired by Salvini and Fabrikant (2016), we chose the *graph embedder* layout (GEM) in NWB to depict the network visualizations (Frick et al., 1995). A disadvantage of the GEM layout is that the *distances* between toponyms in the network visualization are not representative of the *strength of the relationship*, because the algorithm only interprets the presence or absence of a relationship. However, this can be compensated by the visual variables *size* and *color value* as shown in Figure 28. An advantage of GEM, particularly if working with large networks, is that the GEM layout produces graphs which are aesthetically pleasing, as it minimizes *edge crossings* in the network. Details regarding the GEM layout are described by Frick et al. (1995).

In this subsection, we have illustrated how we computed and visualized the spatio-temporal data in the HDS in network visualizations. In the following subsection, we focus on spatializing and visualizing thematic data in the HDS.

4.2.2 Thematic relationships between HDS articles

Participants of a *focus group meeting* (i.e., historians) defined several requirements regarding the visualization of information in the HDS, which are listed in *Subsection 5.3.1*. For this section, the requirement to incorporate thematic information in interactive web interfaces in order to access the HDS content from a thematic point of view is relevant. We chose the *self-organizing map* (SOM) technique to address this requirement as it depicts data elements on a map and highlights the semantic relatedness of data items in the map by placing related data items in similar regions, as introduced in *Section 2.2*. Therefore, identifying and analyzing clusters of semantically similar data items is supported by a SOM. We found this particularly relevant to our project, as we expected that an interactive SOM (see *Subsection 5.2.2*) might help our target users search for interesting themes and articles on the map and identify semantically similar articles more quickly. In addition, SOMs were found to perform very well on large data sets, as explained in *Section 2.2*. Therefore, an interactive SOM might provide efficient access to thematic information, and thus fulfilling the thematic information access requirement of our target users (see *Subsection 5.3.1*). In order to test this assumption, we analyzed if our target users understand and how well they perform tasks with a SOM, which is outlined in *Subsection 5.3.3*.

The computation and visualization of the SOM, as described in this subsection, was inspired by Salvini (2012). The methodological approach is illustrated schematically in Figure 29. The input *thematic database* consists of an *article-topic matrix* with 3,067 *thematic contributions* and their probability distribution over 30 *topics* (see *Subsection 4.1.3*). Several layout and clustering algorithms were applied to this database in order to compute a *clustered SOM cartogram*.

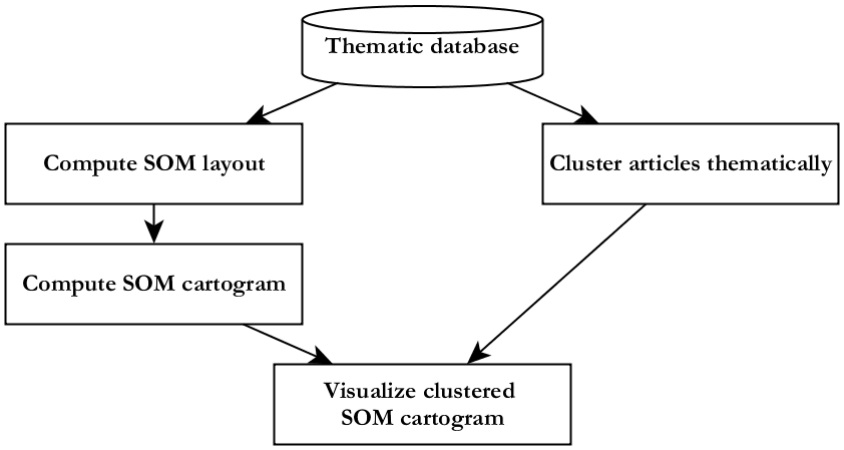


Figure 29: Generalized *self-organizing map* approach for thematic data.

Compute SOM layout

First, the *article-topic matrix* was imported to *SOM Analyst*[59] in *ArcGIS Desktop 10.1*[60]. *SOM Analyst*, developed by Lacayo-Emery (2011), is a toolbox to process and transform input data to a SOM layout. The selection of an appropriate number of neurons was based on Skupin and Esperbé (2011: 293), who suggest defining for each input vector an individual neuron to develop detailed structures among individual input vectors. Therefore, we decided upon a 55*55 neurons SOM (i.e., 3,025 neurons), which creates a similar number of neurons compared to the number of input vectors (i.e., 3,067 articles). We set *hexagonal* neurons as *topology type* which is more frequently used in *information visualization* than the alternative option of *squared* neuron shapes (Skupin and Agarwal, 2008: 8). To create an initial SOM, we chose to assign random values in the *SOM Analyst*, as suggested by Skupin and Agarwal (2008: 8), in order to force a true *self-organization* as opposed to assigning weights according to a *linear estimate* (e.g., *principal components* in training data). In the first training stage, the SOM was trained with 30,000 runs and a neighborhood radius of 55 to establish broad, global structures (Skupin and Esperbé, 2011: 293). In the second training stage, 300,000 runs and a neighborhood radius of 7 were applied to elaborate regional and local structures in the SOM (Skupin and Esperbé, 2011: 293). The *learning rate*, which defines the degree of specialization of regions in the SOM (i.e., high *learning rate* equals high specialization), was set to 0.04 in the first and 0.03 in the second training stage, following the recommendations by Lacayo-Emery (2011: 46-47) and Salvini (2012: 85). The SOM output consists of 30 different *component planes*, which represent the neuron weights of each of the 30 topics in the SOM (Skupin and Agarwal, 2008: 12). Neurons in the SOM are placed close to one another in clustered regions if they share similar topic distributions. We then projected the 3,067 *thematic contributions* articles onto these *component planes* by applying the *best matching unit* (BMU) approach (Skupin and Agarwal, 2008: 13). Thus, each article is placed onto the neuron in the SOM which best matches with its input vector, as depicted in Figure 30 (left). Therefore, HDS articles which are

[59] SOM Analyst: https://github.com/mlacayoemery/somanalyst (accessed July 2016)
[60] ArcGIS Desktop: http://www.esri.com/software/arcgis/arcgis-for-desktop (accessed July 2016)

similar in their topic distribution are placed onto the same or neighboring neurons, and dissimilar articles are placed in neurons which are distant from one another, which corresponds to the empirically tested *distance-similarity metaphor* in region-display spatializations (Fabrikant et al., 2006). The *component planes* and the BMU were stored as separate *shapefiles* in ArcGIS.

Compute SOM cartogram

In the next step, we addressed a conceptual limitation of traditional SOMs: although the distance of neighboring neurons is similar across the SOM, the similarity of neighboring neurons differs across the SOM. Therefore, we decided to distort the SOM layout in order to adapt the distances to the similarities, as suggested by Salvini (2012) and Bruggmann et al. (2013). The similarity of neighboring neurons is described by the *U-matrix*, which is an output of *SOM Analyst* and contains a value for the similarity of each neuron to its neighboring neurons (Ultsch, 1993: 310). The lower the values in the *U-matrix*, the more similar neighboring neurons are (Ultsch, 1993: 310). Following the *distance-similarity principle* (Fabrikant et al., 2006), thematically less related neurons (i.e., high *U-matrix* values) were depicted larger, whereas thematically more related neurons (i.e., low *U-matrix* values) were visualized smaller in size by applying the cartogram technique. In order to compute the cartogram, we transferred the SOM *component planes* to *Scape Toad 1.1*[61] and used the *U-matrix* to distort the neurons in the *component planes*. The output of this process is illustrated schematically in Figure 30.

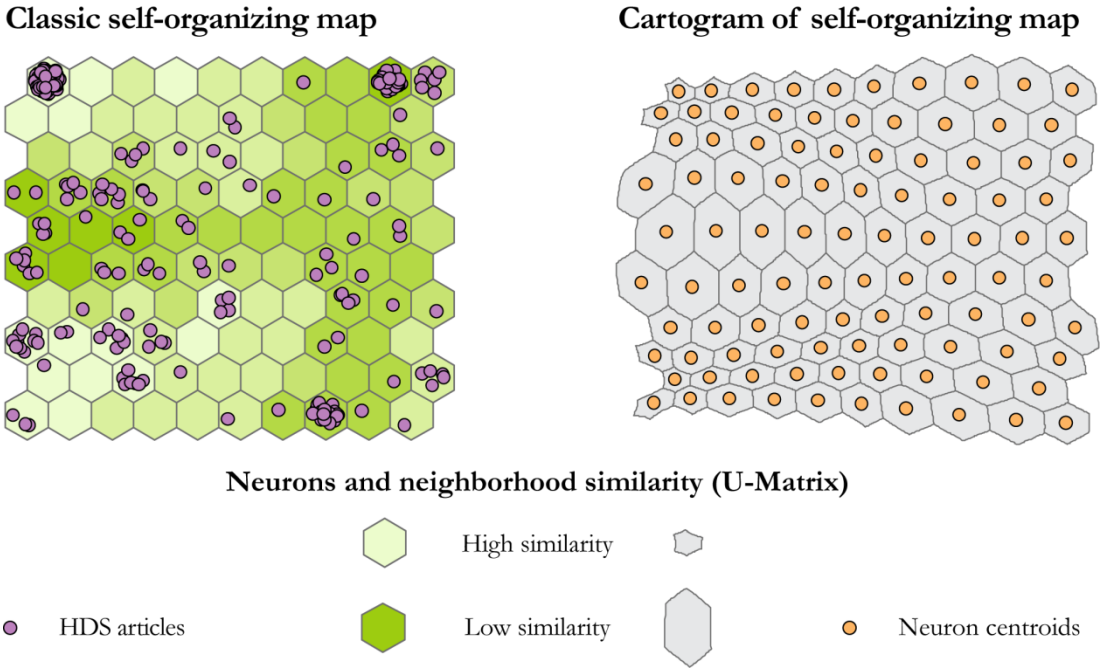


Figure 30: Classic and cartogram of a *self-organizing map* (Bruggmann et al., 2013).

[61] Scape Toad: http://scapetoad.choros.ch/ (accessed July 2016)

Cluster articles thematically

In the next step, we clustered the 3,067 HDS *thematic contributions* articles thematically. We decided to cluster the articles in order to highlight regions of thematically similar articles in the *self-organizing map* (SOM) and to facilitate target users to get an overview of themes which are present in the SOM in the interactive web interface (see *Subsection 5.2.2*). For this reason, we first calculated *cosine similarity* values (i.e., angle between vectors) for all possible article pairs based on the topic distributions over articles by applying the *pairwise distances* method to the *article-topic matrix* using the Python *scikit-learn package*[62]. Different similarity measures have been studied by Ellis et al. (1993: 145) for *text retrieval* tasks, and they propose the *cosine similarity* to measure the degree of similarity between objects in text retrieval systems. The computed *cosine similarity* scores were then transferred to *MySQL Workbench*. Following Salvini's (2012: 110-17) suggestion, we stored the 15 strongest relationships for each article in a *text file* and subsequently transformed the *text file* to a *.net file* in *txt2Pajek*. Then, we computed the *Blondel community detection algorithm* in NWB, which was introduced in *Subsection 4.2.1*. As a result, we identified 28 *themes* (i.e., clusters) in the HDS *thematic contributions* articles.

Visualize clustered SOM cartogram

In the final step, we visualized the SOM cartogram and the BMU, including their membership to *themes* in ArcGIS, following the *spatialization framework*. Two different ways of representing the clustered SOM cartogram were chosen (i.e., *detail SOM* and *overview SOM*) and are illustrated schematically in Figure 31. The combined *detail* and *overview SOM* view follows Shneiderman's (1996) *visual information-seeking mantra*: "*overview* first, *zoom* and *filter*, then *details-on-demand*". Therefore, we planned to present an overview of the SOM (i.e., *overview first*) in an interactive SOM interface to target users, which should assist them in obtaining an overview of the *themes* present in the SOM. By zooming in to a specific *theme*, the *detail SOM* (i.e., *details-on-demand*) and thus single *thematic contributions* articles are presented to them. The combination of the *detail* and the *overview SOM* in an interactive web interface, including interaction functionalities such as *zooming*, is presented in *Subsection 4.3.2*.

First, *semantic generalization* was performed: single articles (i.e., BMU) were interpreted as *loci*, neurons and regions of thematically similar articles as *aggregates*, and borders of neurons and regions as *boundaries*. Second, *geometric generalization* was applied: the geometric primitive *point* was assigned to the semantic primitive *locus*, *line* to *boundary*, and *area* to *aggregate*, respectively. Furthermore, the visual variable *color hue* was used to depict the *theme*, and the visual variable *size* was employed to differentiate *neuron boundaries* from *article area boundaries* (i.e., *themes*).

---

[62] *Pairwise distances* method in the *scikit-learn* package in Python: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.pairwise_distances.html (accessed July 2016)

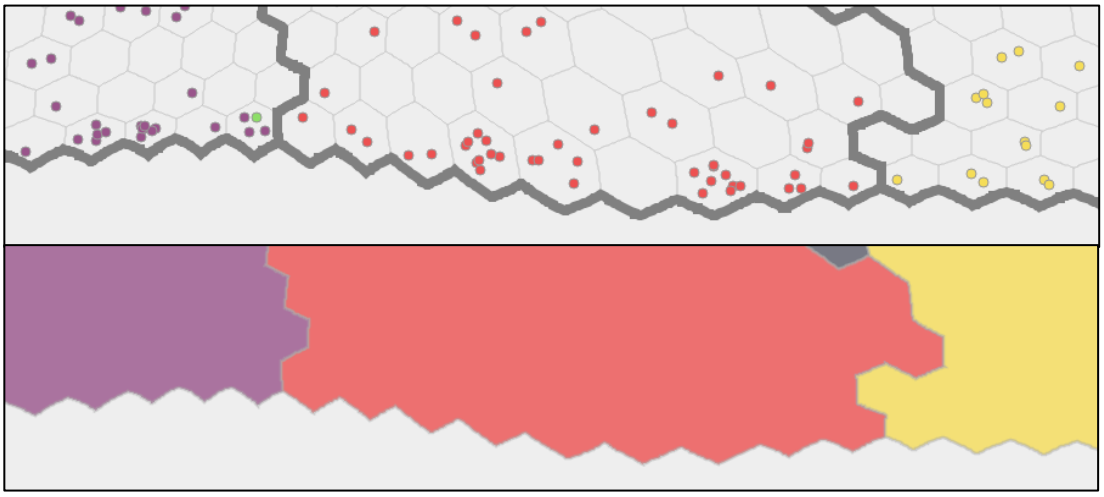


Figure 31: *Detail SOM* (above) and *overview SOM* (below).

The *detail SOM* and *overview SOM* in Figure 31 depict the same information, but are different from a *cartographic generalization* point of view: the *overview SOM* was generalized by resolving the boundaries of neighboring neurons, which contain articles of the same *theme* in ArcGIS. If neurons from more than one *theme* were placed onto one neuron, the *theme* with the highest frequency of articles in the neuron was assigned to the neuron. The boundaries of the *article cluster areas* are also depicted in the *detail SOM*, but are illustrated with a larger border in order to become distinguishable from the neuron boundaries.

To summarize, we have illustrated the transformation and reorganization of spatio-temporal and thematic information in the HDS and the visualization of this information in spatialized displays (i.e., *network visualization*, *self-organizing map*) in this section. In the following section, we present the incorporation of these spatialized displays in interactive web interfaces by applying a user-centered evaluation and design approach, thus outlining how we plan to address *Research Question 3* of this thesis.

4.3 Geovisual analytics

We aim at making the spatialized displays available to the target users of this project (i.e., historians, people interested in *digital humanities*) to support sense-making and the generation of new insights regarding spatial, temporal, and thematic information in the HDS. Motivated by Moretti (2005) and Jockers (2013), we decided to implement interfaces which incorporate coupled *distant* and *close reading* functionalities. We selected *geovisual analytics* for this project as it provides a methodological framework to incorporate spatio-temporal and thematic information in exploratory web interfaces that support the generation of new insights. Furthermore, *geovisual analytics* suggests specific methods to involve target users in the interface design and evaluation process which we found particularly relevant for the development of interfaces which fit the information-seeking needs of our target group. The approach we employed in this

project was inspired particularly by Roth et al. (2015) and Lewis and Rieman (1993), which was explained in *Subsection 2.3.2*. Several parts of this approach have been published in Bruggmann and Fabrikant (2016).

The general workflow for this process is depicted in Figure 32. It follows the iterative user-centered interface design and evaluation approach introduced by Roth et al. (2015), as described in *Subsections 2.3.2* and *2.3.3*. In the left column of Figure 32, several runs of the iterative *user-utility-usability* triangle are schematically illustrated. In the middle column, the geoVA approach applied by Roth et al. (2015) is shown and compared to our approach in the right column. The grey portion in Figure 32 represents the main differences between the two approaches. We involved the designer in a *cognitive walkthrough* which is inspired by Lewis and Rieman (1993) and detailed in *Subsection 4.3.1*.

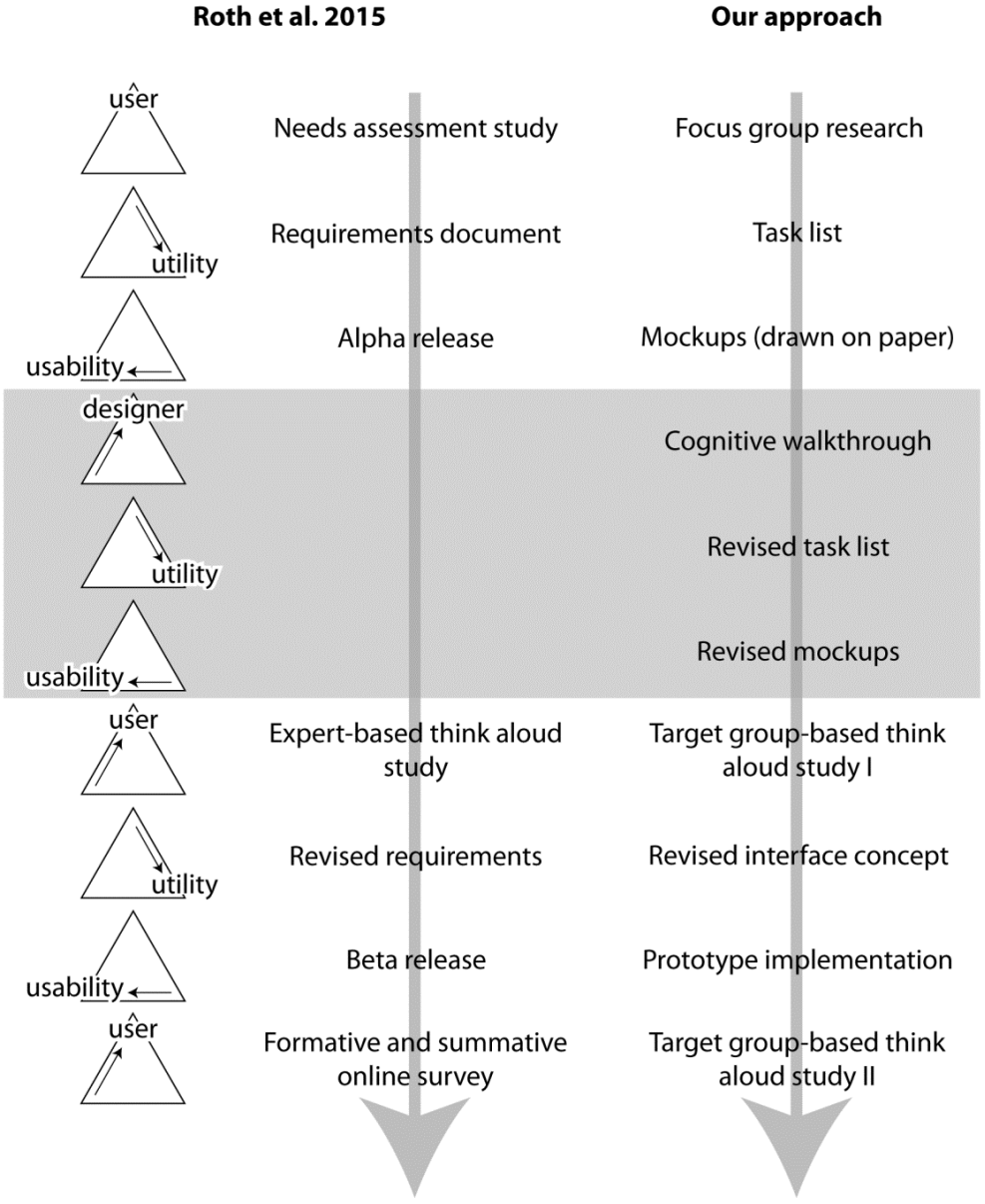


Figure 32: Iterative *geovisual analytics* workflow from Roth et al. (2015) (left and middle column), compared to our own approach (right column) (modified from Bruggmann and Fabrikant, 2016).

We first introduce the empirical evaluation of the developed user interface design in *Subsection 4.3.1*, which covers all but the last two steps of the right column in Figure 32. In *Subsection 4.3.2*, the implementation of two prototype web interfaces is explained, which is the penultimate stage of our approach in this project. The final stage in Figure 32 considers the evaluation of the prototype implementation, and is detailed in *Subsection 4.3.3*. As introduced in *Chapter 3 – Data*, we defined the target users in this project as historians who are interested in new media types and methods in history, people who are interested in *digital humanities*, and those who are interested in interactive interfaces with which to explore the humanities in general.

In the following subsections, only the *participants*, *procedure*, and *methods* to analyze the results of the empirical evaluations are presented. The results of the empirical evaluations, the task lists, and the paper mockups we drew and presented to target users are inputs as well as the results of the empirical evaluations, as demonstrated in Figure 32. Therefore, these parts are presented in the results of this thesis in *Section 5.3*.

4.3.1 Empirical evaluation of user interface design

Focus group research

The goal of the *focus group research* was to discuss our *spatialized networks* approach (see *Subsection 4.2.1*) and our initial design ideas to depict the networks in interactive web interfaces for potential target users of our project (Rubin and Chisnell, 2008: 17). We chose this method as it helps to assess the poorly known needs and expectations of target users at an early stage of the user-centered interface design process (Roth et al., 2015: 276). We followed a four-step approach outlined by Kessler (2000: 45-55) and adapted it to our project needs, as detailed in the following list.

1) **Planning phase**. In the planning stage, we identified the *goals* of the *focus group research*: we were interested in the opinions of target users regarding the spatio-temporal networks we created. In addition, we were interested in their opinion on our idea of interactively visualizing the spatio-temporal networks in web interfaces and if and how they would use them. Lastly, we were interested in input from the target users regarding further ideas on how to depict spatial and temporal data interactively.
   We did not include *self-organizing maps* in the *focus group meeting*, because we only planned to include them in our project after the meeting (i.e., based on the feedback of the *focus group* participants). This is detailed in *Subsection 5.3.1*.

2) **Recruitment of focus group participants**. We contacted potential target users from both inside and outside of the University of Zurich. We recruited five people fitting the target user requirements. Four of them studied *history* as a major subject at university, and one participant has *cultural studies* as a major. Two participants also studied *geography* as a minor. All five participants have graduated from university, know and have used the HDS, and are dealing with Swiss history in their job. Four of the participants indicated that they have no experience with implementing

interactive web visualizations. Our group size is in the lower range of the commonly suggested 4-12 members by the literature (e.g., Tong et al., 2007: 351). With this comparatively small group size, we aimed at forcing in-depth discussions among participants and providing each participant with enough time to provide input to the discussion, as suggested by Onwuegbuzie et al. (2009: 3).

3) **The focus group meeting**. During the first 30 minutes of the meeting, the author of this thesis presented the aim of the project and the *focus group meeting*, the initial results, and some hand-drawn sketches containing web interface ideas. We decided to use hand-drawn sketches instead of a fully implemented tool as we aimed at providing target users the impression that the interface ideas are incomplete and non-definite, and thus that changes are easily possible at this stage, which was suggested for early stages of an iterative interface design process for example by Wong (1992) and Landay and Myers (2001). A 60-minute discussion followed which was moderated by the author of this thesis. The discussion was audio-taped using the pre-installed *Voice Memos* app on an *iPhone 6*[63]. The moderator raised questions, promoted verbal interactions between participants, made sure that the discussions stayed on-topic, and monitored discussion time in order to allocate sufficient time to receive responses to all prepared questions. Most of the questions were open-ended in order to encourage participants to discuss their thoughts in greater detail without constraint. The entire meeting was held in German, as all participants have German as their native language.

4) **Analysis of focus group**. During the *focus group meeting*, requirements for the potential web interfaces and a task list were elaborated (see *Subsection 5.3.1*). After the meeting, the task list was revised.

Roth et al. (2015) applied a similar approach to assess the requirements of their target group, but conducted interviews instead of a *focus group meeting*. The *questions*, the *sketches*, and the *results* of the *focus group meeting* (i.e., *requirements* and *task list*) are presented in *Subsection 5.3.1*.

Cognitive walkthrough

For the next evaluation step, we employed the *cognitive walkthrough* method, which is an evaluation method without users. Not considering target users at this stage allows the designers of a user interface to remove as many problems as possible before target users are involved in the design evaluation process (Lewis and Rieman, 1993: 41). As illustrated in Figure 32, we used the task list generated as an output of the *focus group meeting* to draw mockups of the planned web interfaces on paper. During the *cognitive walkthrough*, these mockups were used to evaluate the planned design of the interfaces. The results of the *cognitive walkthrough* aided our revision of the task list and mockups, as shown in Figure 32. The *tasks*, *mockups*, and *results* of the *cognitive walkthrough* are presented in *Subsection 5.3.1*.

[63] Apple's iPhone: http://www.apple.com/iphone/ (accessed October 2016)

For each task on the list elaborated in the *focus group research*, we constructed a series of actions which would need to be followed in order to solve the respective information-seeking tasks. Actions are, for example, clicks on hyperlinks or buttons, clicks on a drop down menu to select a menu option, or to navigate to a website. Paper mockups were developed for each action, and thus the state of the interface before and after the execution of actions was sketched. In the analysis stage of the *cognitive walkthrough*, the designer (i.e., the author of this thesis) goes through all the tasks and attempts to decide, for each action, whether the target users might choose the correct action or not. All decisions have to be documented, and *success* and *failure stories* also have to be reported on (Lewis and Wharton, 1997: 722-23). A *success story* is only identified if all of the following questions can be answered with *yes* by the designer (Lewis and Wharton, 1997: 722).

- Will the user be trying to achieve the right effect?
- Will the user notice that the correct action is available?
- Will the user associate the correct action with the desired effect?
- If the correct action is performed, will the user see that progress is being made?

If one or more of the aforementioned questions are answered with *no*, a *failure story* is identified. Hence, a potential problem with the web interface design is identified which results in a revised set of task list and respective mockups, as depicted in Figure 32.

Target group-based think aloud study I

In the next step, we conducted an initial *think aloud study* based on the revised task list and respective mockups of the *cognitive walkthrough* (see *Subsection 5.3.1*), as illustrated in Figure 32. The basic idea of a *think aloud study* is to ask participants to solve a task and to comment on their thoughts and decisions while performing this task in order to identify potential interface design problems (Lewis and Rieman, 1993: 83). Here, we present information regarding our *participants*, the *procedure*, and how we analyzed the *results*. The *tasks*, *mockups*, and *results* of this *think aloud study* are presented in *Subsection 5.3.1*.

- **Participants**. We recruited five participants following Nielsen's (1994) suggestion for a *think aloud study*. Following Lewis and Rieman (1993), but in contrast to Roth et al. (2015), we recruited members of our target group as participants instead of design experts, as illustrated in Figure 32. We selected participants who have at least some experience with interactive web interfaces, as we wished to get specific feedback on possible interface design issues. Three of the five participants have an educational background in geography, are experienced with the design of interactive web interfaces, and have an interest in Swiss history, while two participants are historians. One of the historians has little and the second some experience with interactive web interfaces.

- **Procedure**. The participants were invited to take part in individual sessions which were videotaped using the pre-installed *Camera* app on an *iPhone 6*. The sessions lasted approximately 60 minutes. The author of this thesis moderated the sessions. We conducted an initial *pilot study* to test the *study design* and *procedure*, then adjusted these accordingly based on feedback. Prior to beginning the videotaped portion of the main *think aloud study* the participants signed a *consent form*. Furthermore, they were asked to read a printed handout which detailed the aim of the study and some general information regarding the research project. In addition, they were introduced to the *think aloud* procedure: we asked them to comment their actions, decisions, and any potential issues they might face during the study. They were also instructed to speak in their native language, which was German for all participants. After having solved a practice task, the videotaped portion of the study began. The tasks were given to participants on printed handouts. Participants solved the tasks with mockups that illustrated the state of the interface, depending on the interactions of the participants with the interface. The moderator did not respond to questions and participants were not given any help during the study. The moderator only provided help if participants were completely lost. If participants decided on an action for which no mockup had been prepared, a hand-drawn sketch with a dialog box containing an error message was presented to them which stated that the chosen action is not available in the current version of the interface, and that they should please choose another action. The moderator kept notes during the sessions and asked participants to comment on specific situations that had occurred (e.g., participants seemed completely lost), about the interface in general, and ideas for potential improvement in a short debriefing session. After completing the session, the participants were given a small present to thank them for their participation.
  In Figure 33, the experimental setup with the participant, moderator, video camera tripod, video camera (i.e., an *iPhone 6*), and videotaped area with hand-drawn mockups is shown.

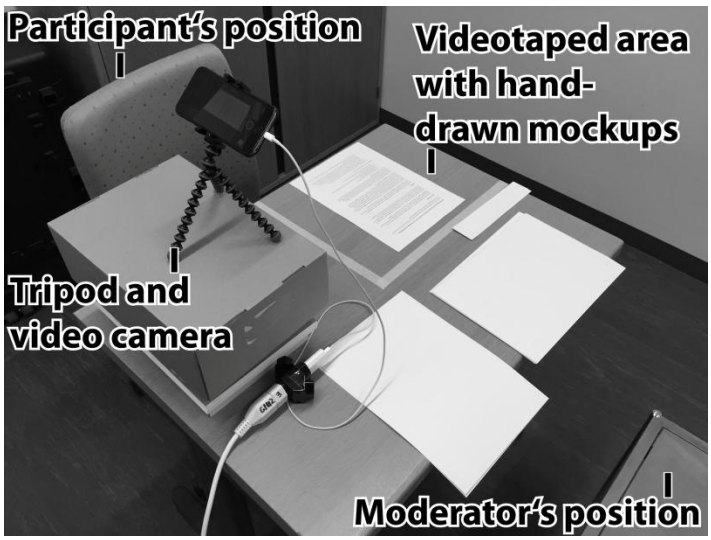


Figure 33: Experimental setup of *think aloud study I* (Bruggmann and Fabrikant, 2016: 9).

- **Data analysis**. After the *think aloud sessions*, the author of this thesis studied the video tapes, and issues with the interface concept were identified by comparing the *action sequence* of the participants with the expected *action sequence* defined in the *cognitive walkthrough*. Issues of the interface were judged according to their *importance* and the *difficulty* to fix them (Lewis and Rieman, 1993: 86). The *importance* judgment was based on the potential costs of the problems to the users (e.g., in time, aggravation) and the likely proportion of users which might face similar problems. *Highly important* and *easy to fix* issues were considered to be more relevant, compared to issues with a *low importance* and which are *very difficult* to be fixed (Lewis and Rieman, 1993: 86). All results are detailed in *Subsection 5.3.1*.

Based on the *think aloud study* results, we developed a *revised interface concept* which served as an input for the *prototype implementation* process. The *revised interface concept* consists of a description of the planned web interfaces, including planned interface and interaction elements, as described in *Subsection 5.3.2*.

4.3.2 Prototype implementation

The penultimate stage (see Figure 32) considers the implementation of prototype web interfaces. We implemented two interactive web interfaces: a *spatialized network interface* and a *thematic landscape*, visualized in a SOM. Both visualizations incorporate combined *distant* and *close reading* functionalities (Moretti, 2005, Jockers, 2013) and are designed based on Shneiderman's (1996) *visual information-seeking mantra*: "*overview* first, *zoom* and *filter*, then *details-on-demand*". The interface concept was guided by the user-centered interface design evaluations described in the previous subsection (see the results of the evaluations in *Subsection 5.3.1*). The aim of the prototype implementations is to provide information seekers interested in Swiss history with exploratory and interactive web interfaces to support sense-making and the generation of new insights regarding spatio-temporal and thematic information and interconnections in the HDS. Both interfaces are accessible on the author's personal website (as from spring 2017: http://www.geo.uzh.ch/~abruggma/index.html). On this website, further information regarding the project is available. All contents on the website and in the two web interfaces are only available in German, as we used the website to test the interfaces with our target users, all of which have German as their native language.

In the following sections, the technical implementation of prototypes is explained. The web interfaces are depicted and described in detail in *Subsection 5.3.2*.

Spatialized network interface

We included the *spatialized network* visualizations (see *Subsection 4.2.1*) in a web page. Due to feedback from target users in the empirical evaluations (see *Subsection 5.3.1*), we decided to implement a dynamic and interactive web interface for two spatial and three temporal scales for our prototype implementation. For temporal scales, we chose the 18th, 19th, and 20th centuries because these three centuries are most covered in the HDS and we therefore expected that the most complete network of Swiss toponyms would

be drawn (see *Subsection 5.1.2*). In addition to the network toponyms from throughout Switzerland, we decided on the *Canton of Zurich* to serve as an additional spatial scale because the major structures on the territory of today's *Canton of Zurich* have remained quite stable in the past when compared to other cantons: from the mid of the 15th century to 1798, the city of Zurich owned the territory in the area of today's *Canton of Zurich*. After 1798-1803, the *Canton of Zurich* was founded as a part of the *Helvetic Republic*, and since 1803, the *Canton of Zurich* is part of the *Swiss Confederation* (Horisberger et al., 2015). Another reason for choosing the *Canton of Zurich* was that most participants in our empirical evaluations possess detailed knowledge regarding the *Canton of Zurich*, as they work in and have additional ties to Zurich, and some of them deal with the history of the *Canton of Zurich* as a part of their job. Both the architecture of the interface and pseudo code to incorporate the elements on a website are illustrated in a simplified manner in Figure 34.

As we wished to create an interface which is accessible for our target users online, we chose *Hypertext Markup Language* (HTML)[64], the standard markup language for creating web pages, to embed our web interface. We chose *Scalable Vector Graphics* (SVG)[65] as a language to describe network visualizations because SVG allows us to control the location of each single object (e.g., network nodes and edges) in the visualization, since coordinates for all objects must be defined for each object. Therefore, we were able to place all objects (i.e., network nodes and edges, interaction elements) at the exact locations we wished them to be in order to comply with the mockups of the *think aloud study* (see *Subsection 5.3.1*). Furthermore, SVG is vector-based, meaning SVG graphics are freely scalable which serves as an additional advantage because they easily adapt to various screen sizes. Interactivity of HTML and SVG elements was implemented with *JavaScript*[66], which was chosen as it is a core technology in the creation of dynamic web sites and is supported by all modern browsers without an additional plug-in. *Cascading Style Sheets* (CSS)[67] was chosen to style elements (e.g., color hue of nodes) in the HTML and SVG codes. The primary advantage of CSS is that it can lay out multiple objects at once (e.g., all nodes of a specific class are red) and thus reduces code size, allowing web pages to load faster when compared to a system in which each element is assigned a separate styling element.

To the right of the pseudo code in Figure 34, print screens of the respective elements are illustrated. The entire interface is depicted and described in *Subsection 5.3.2*. In the top right corner of the boxes in Figure 34, the scripts which were applied to the different blocks of code are depicted (i.e., *CSS*, *JavaScript*). In the bottom right corner, the file type is indicated (i.e., *HTML*, *SVG*).

[64] Hypertext Markup Language (HTML): https://www.w3.org/html/ (accessed April 2016)
[65] SVG: https://www.w3.org/TR/SVG11/ (accessed July 2016)
[66] JavaScript: http://www.w3schools.com/js/ (accessed July 2016)
[67] Styling HTML with CSS: http://www.w3schools.com/html/html_css.asp (accessed July 2016)

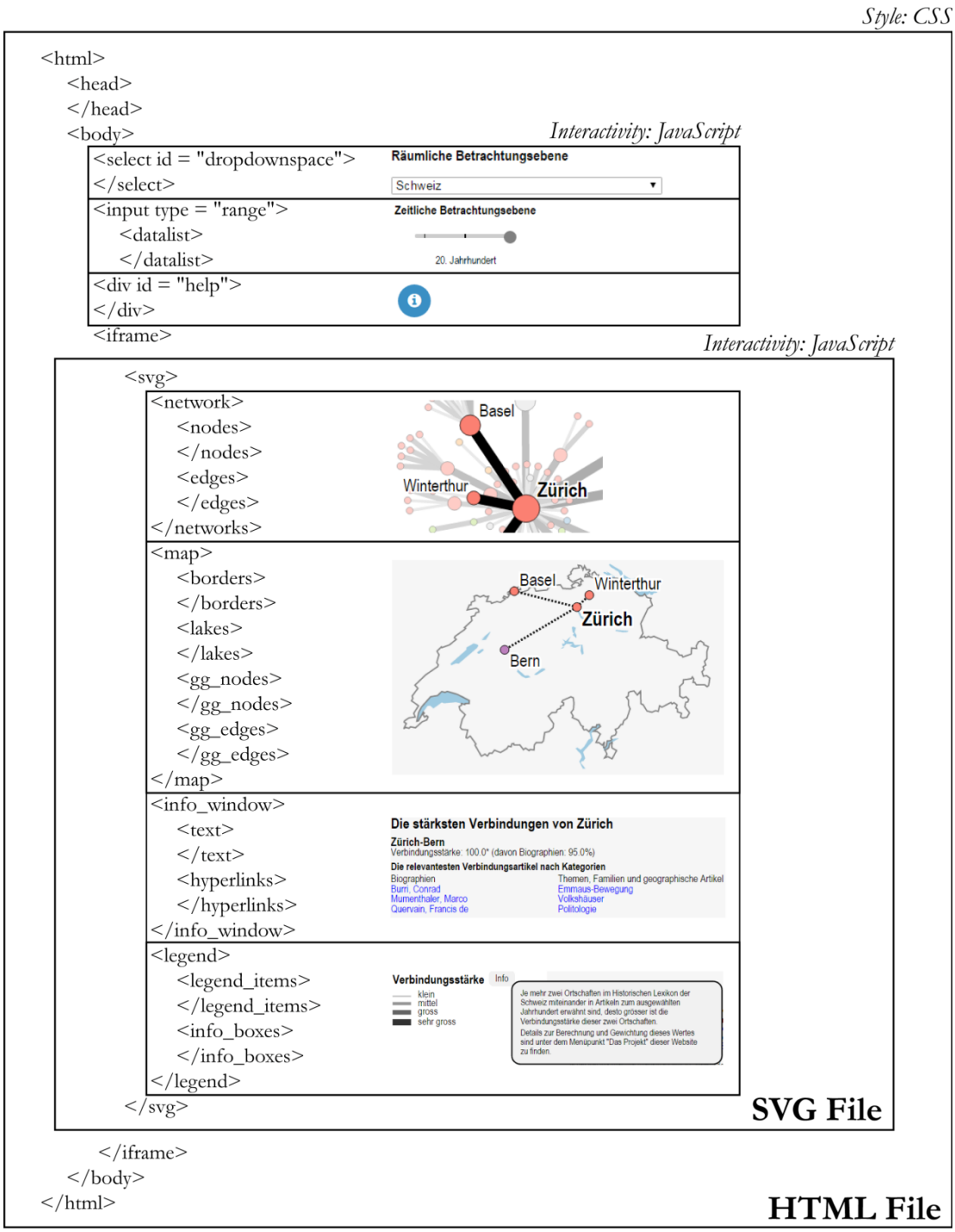


Figure 34: Architecture and pseudo code of the *spatialized networks interface*.

The HTML file contains the contents of the website. Within the HTML file, elements are styled by a CSS file, meaning that *text fonts*, *color hues*, *color values*, the *size* of lines or points, and other *style elements* were applied to elements in the HTML and SVG files. In the *body* part of the HTML file there is a *select*, an *input*, and a *div* element. The *div* element represents a button. If a user clicks this button, help and information regarding the web interface is provided in a pop-up window. The *select* element creates a drop-down menu to select a *spatial scale*, while the input type *range* creates a time slider to

select a *temporal scale*. The drop-down menu was chosen as it is a common interaction element in web design to select an item and we thus expected our target users to understand how this element works. The choice for the *time slider* was based on participants' feedback in the *focus group meeting* (see *Subsection 5.3.1*). Depending on the selected *spatial* and *temporal scale*, the respective SVG file is embedded in the website by an *iframe* element. The *interactivity* and loading of the correct SVG file is controlled by a *JavaScript* file. The SVG file, which is embedded as an *iframe*, contains several groups of elements: the *network* group contains all *nodes*, *edges*, and the respective *labels* in the network visualization. Each node and line has a *network coordinate*, and the *style elements* are applied by the CSS file. The *map* group contains the coordinates of the *border* of Switzerland and of the *Canton of Zurich* as well as the coordinates of the largest lakes in Switzerland. Furthermore, all geographic coordinates of the nodes and edges (i.e., *gg_nodes* and *gg_edges* in Figure 34) are stored in the *map* group in order to depict the spatio-temporal relationships of toponyms geographically. The *info_window* group contains information regarding articles in which two toponyms co-occur, which is explained in detail in *Subsection 5.3.2*. The *legend* group contains all elements which describe the network and the map: for example, the *strength of the relationship* (= *Verbindungsstärke*). Additionally, *information boxes* (= *Info* in Figure 34) describe various elements of the interface that are part of the *legend* group. The interactivity within the SVG file is controlled by a *JavaScript* file. For example, if a user moves over a node or an edge with the mouse cursor in the network visualization, spatio-temporal relationships in the network visualization and in the map are visually highlighted.

The *node* and *edge coordinates* of the network visualization, as well as the *Blondel community membership* of toponyms, were exported from NWB, and the *geographical coordinates* of the toponyms from the *SwissNames* data set. All information related to *spatio-temporal relations* were exported from the spatial and temporal *MySQL Workbench* database tables. All information was imported in *Microsoft Excel* and combined in order to create separate SVG files for each century and spatial scale manually by applying *Excel formulas* (e.g., VLOOKUP). The *geographical coordinates* of the borders and lakes in the map were retrieved from the *ThemaKart* data set of the *Swiss Federal Statistical Office*[68].

The functionalities of the interface and the interplay of the functionalities are detailed in *Subsection 5.3.2*.

Thematic landscape

We implemented the *thematic landscape* in *ArcGIS Online*[69], which is a *cloud-based* mapping platform that allows the export of data from *ArcGIS desktop* to *ArcGIS Online*, and to optimize the design in an *online mapping interface*. This was particularly useful for our project as we created the *self-organizing map* in *ArcGIS desktop* and were therefore able to simply integrate the data into the online platform. Furthermore, it is possible to

[68] ThemaKart on the website of the Swiss Federal Statistical Office: http://www.bfs.admin.ch/bfs/portal/de/index/regionen/thematische_karten/01/02.html (accessed July 2016)

[69] ArcGIS Online: http://www.esri.com/software/arcgis/arcgisonline/ (accessed July 2016)

implement an *interactive web application* based on the data in the *online mapping interface* without coding as it provides many built-in *widgets* (e.g., zoom slider). Most interaction elements (i.e., *widgets*) which we wished to include in the *thematic landscape* were accessible in *ArcGIS Online*, which was of great benefit to the development process, as detailed in *Section 5.3.2*.

We first transferred the *detail* and the *overview SOM* (see *Subsection 4.2.2*) as *shapefiles* to *ArcGIS Online* and adapted the design in the *online mapping interface*. In the *detail SOM*, we created labels for the articles, which are displayed in the top right corner of the article points in the SOM. Additionally, pop-up windows for the articles in the *detail SOM* were implemented, which open if an article point in the SOM is clicked. In these pop-up windows, the name and a hyperlink to the selected article as well as the 10 thematically most related articles (including a hyperlink to these articles) are displayed. The 10 thematically most related articles were assessed by calculating the *cosine similarity* values of all article pairs based on the topic distributions over articles and selecting the 10 strongest relationships of each article. The calculation is equal to the procedure presented in *Subsection 4.2.2*.

The *overview SOM* was only slightly adapted in *ArcGIS Online*: we visualized the number of articles per *theme* in the *overview SOM*. Furthermore, we defined *descriptive terms* for all *themes* and displayed them in the *overview SOM*. In order to locate descriptive terms, we presented the *detail SOM* with the labeled articles to three people and asked them to define the most descriptive terms for each cluster of articles. The participants were also allowed to access the articles on the e-HDS. If the three people did not agree on descriptive terms for the clusters, the author of this thesis chose a descriptive term instead.

The *detail SOM* and the *overview SOM* were then transferred to the *Web AppBuilder for ArcGIS*[70] in *ArcGIS Online*. The *Web AppBuilder* provides functionalities to implement applications which run on any device (i.e., *desktop computer*, *tablet*, *smartphone*). No coding is required, as previously mentioned. We decided to provide *zooming functionalities* and an *article search widget*. The zooming widget allows users to zoom in to the *detail SOM* or zoom out to the *overview SOM*. Additionally, we incorporated an *information button*. By clicking on this button, help and information about the application is provided in a pop-up window.

In this subsection, we have detailed the implementation of the prototype web interfaces which are presented and discussed in detail in the following chapter of this thesis. In the following subsection, we explain how we evaluated the two web interfaces.

4.3.3 Empirical evaluation of prototype

In order to empirically evaluate the prototype web interfaces presented in the previous subsection, we employed a second *think aloud study*, which is the final stage of our approach, as shown in Figure 32. This diverges from Roth et al. (2015), who conducted

[70] Web AppBuilder for ArcGIS: http://doc.arcgis.com/en/web-appbuilder/ (accessed July 2016)

formative and *summative online surveys*, as illustrated in Figure 32. We decided to conduct *think aloud studies* instead of *online surveys*, as we wished to demonstrate the functionalities of the interfaces to participants before they began the *think aloud study* and answered questions that participants might have had before they began with the study. In addition, we wished to log participants' interactions and audio-tape their comments during the study. Furthermore, conducting a *think aloud study* instead of an *online survey* allowed us to discuss potential issues that participants faced during the study in a debriefing session. The procedure of the *think aloud study* is similar to the first *think aloud study* (see *Subsection 4.3.1*). However, compared to the first *think aloud study*, we were interested in evaluating the *utility* and *usability* of the prototype implementations instead of evaluating the *design* of the interfaces. In particular, we were interested in evaluating the *process of gaining insights*, as introduced in *Subsection 2.3.2*. The approach we chose was particularly inspired by Nelson et al. (2015). However, as a difference to our approach and similar to Roth et al. (2015), Nelson et al. (2015) conducted an *online survey* instead of a *think aloud study*.

- **Participants**. We invited the five participants from our target group who already participated in the *focus group meeting*, and all agreed to return to participate in the study. We chose this approach because they were already familiar with the project and were interested to see the prototype implementation. A detailed description of the participants and their background is presented in the *focus group research* section of *Subsection 4.3.1*.

- **Procedure**. We first conducted three *pilot studies* to test the *study design* and *procedure*, and adjusted these accordingly based on feedback. One week before the study, we informed participants about the planned procedure of the *think aloud study* and sent them an *instruction manual*. The *instruction manual* contained information on how to navigate the website, and how to use the web interfaces (i.e., *spatialized network interface* and *thematic landscape*). We provided participants with access to the password-protected website containing the two interfaces and asked them to familiarize themselves with the basic functionalities of the website and interfaces prior to the study.

  The participants took part in individual sessions which lasted approximately 90 minutes. The author of this thesis moderated the sessions. The study language was German, as all participants had German as their native language. Before starting the *think aloud* portion of the study, participants were seated in front of a monitor (see Figure 35) and signed a *consent form*. Next, the participants were asked to read a handout which detailed the aim of the study and the *think aloud* procedure: we asked participants to comment their actions, decisions, and potential issues they might face during the study. Participants then solved a practice task to get comfortable with the *think aloud* procedure. Next, the moderator explained the functionalities of the *spatialized network interface*, and answered open questions to participants. The first task was then presented (the task is detailed in *Subsection 5.3.3*). Participants were given a maximum of 40 minutes to interact with the *spatialized network interface* in order to solve the task. The moderator did not

respond to questions, and participants were not given any help while solving the task. The moderator only provided help if participants were completely lost.
After 40 minutes had passed, the moderator demonstrated the functionalities of the *thematic landscape* to participants, and answered open questions. Then, the second task was presented (the task is shown in *Subsection 5.3.3*). Participants were then given a maximum of 15 minutes to interact with the *thematic landscape* interface in order to solve this task. Same as for the first task, the moderator did not respond to questions and participants were not given any help while they were solving the task. The moderator only provided help, if participants were completely lost. Participants' interactions with the interfaces were recorded automatically and participants were audio-taped throughout the *think aloud study* with the *ShareX 10.9.1*[71] software installed on the test laptop. After participants completed both tasks, they were given a small present to thank them for their participation.
The experimental setup is illustrated in Figure 35. The participant on the right is seated in front of the monitor. On the left, a test laptop is shown. On this test laptop, we ran the *ShareX* software to log the interactions of the participants with the interfaces and to audio-tape the *think aloud*. This test laptop was connected to the monitor in the middle of Figure 35. On this monitor, the website containing the two interfaces which were evaluated in the *think aloud study* was presented to the participants. In Figure 35, the *thematic landscape* interface is displayed on the monitor (see *Subsection 5.3.2*). A keyboard and a mouse were available to the participants to navigate the website and to interact with the interfaces.

Figure 35: The experimental setup of *think aloud study II*.

After the first and second tasks, participants were asked to fill in a paper and pencil version of the *System Usability Scale* (SUS) questionnaire (Brooke, 1996). The SUS assesses the *global usability* of a system with general measures (Brooke, 1996). It consists of ten questions aimed to assess the *effectiveness* (i.e., ability of performing a

[71] ShareX: https://getsharex.com/ (accessed July 2016)

task), *efficiency* (i.e., the level of resources used to perform a task), and *satisfaction* of the users with the system, and is measured on a five point *Likert scale* (Brooke, 1996). In addition, participants were asked about how satisfied they are with the results they obtained, how confident they are that they successfully completed the tasks, and how relevant they think their insights are regarding the history of Switzerland. Furthermore, the participants were invited to provide further open feedback regarding the user interfaces.

- **Data analysis**. After the *think aloud studies*, the author of this thesis studied the *participants' interaction* and *audio records*, then listed and summarized insights regarding the history of Switzerland that participants gained during the study. For the first task (i.e., *spatialized network interface* task), the insights were ranked in the following categories as suggested by North (2006: 20), on a five point *Likert scale*.

  - **Complexity**. An insight is rated high for complexity if it involves a large amount of the given data in a synergistic way.

  - **Depth**. An insight has a high score for depth if the insight includes many steps, and accumulates on itself over time.

  - **Unexpectedness**. An insight is rated high for unexpectedness if it is unpredictable, serendipitous, and creative.

  - **Relevance**. An insight is rated high for relevance if it is deeply embedded in the data domain and connects data to existing domain knowledge.

  The first three categories were evaluated by the moderator of the *think aloud study*. As suggested by North (2006), for the *relevance* category, an expert historian was involved for rating the insights. We excluded the *quality* category (e.g., the *accuracy* of an insight) from North's (2006) list, but invited the expert historian to provide further comments on the accuracy of an insight.
  For the second task (i.e., *thematic landscape* task), answers which were provided by the participants to the task were graphically compared to one another, as detailed in *Subsection 5.3.3*.

In this chapter, we have illustrated the user-centered interface design and evaluation process applied in our project and the prototype implementations we developed based on the feedback received from our target users in the *focus group* and the *think aloud sessions*. In the following chapter, we illustrate the results obtained by applying the methods illustrated in this chapter and provide answers to the three research questions of this thesis.

5 Results

In this chapter, we present the results of this thesis. In *Section 5.1*, we illustrate the spatial, temporal, and thematic information extracted from the HDS by applying *geographic information retrieval* methods. We transformed and visualized the retrieved spatio-temporal and thematic information in spatialized displays (i.e., *network visualizations*, *thematic landscape*), which are presented in *Section 5.2*. In the next step, we incorporated the spatialized displays in exploratory and interactive web interfaces, involving target users of our project in the interface design and evaluation process. Therefore, we present the evaluation results of the user studies and how we incorporated the evaluation results into the interface concept in *Section 5.3*. Furthermore, we introduce prototype implementations of the *spatialized network interface* and the *thematic landscape*.

5.1 Geographic information retrieval

In this section, we analyze the spatial, temporal, and thematic data retrieved from the HDS articles. The four HDS article categories (i.e., *biographies*, *geographical entities*, *thematic contributions*, and *families*) differ substantially regarding the amount of spatial and temporal information they contain; therefore, these categories contribute differently to the spatio-temporal relationships we depict in spatialized displays (see *Section 5.2*). We illustrate differences regarding the amount of spatial and temporal information in articles of these categories in *Subsections 5.1.1* and *5.1.2*, respectively. Furthermore, we illustrate the spatial distribution of toponyms to illustrate the coverage of different regions of Switzerland in this project. In addition, we present the number of temporal references by *centuries*, as we focus on temporal information at the *centuries* level in this project. In *Subsection 5.3.3*, the thematic information retrieved from the HDS articles by applying the *topic modeling* approach is shown in order to illustrate the thematic diversity covered by HDS articles.

5.1.1 Spatial data

We retrieved a total of 322,179 toponyms from the 36,188 HDS articles, of which 16,489 toponyms are unique. In other words, a toponym is mentioned 19.5 times in the HDS corpus on average. We determined that 97.9 % of all articles contain at least one

toponym. As outlined in *Subsection 4.2.1*, we only considered the *settlements* feature type of the *SwissNames* gazetteer in this project. *Settlements* account for 264,703 toponyms and thus cover 82% of all retrieved toponyms. Here, we use toponym as a synonym for the *settlements* feature type.

Toponym occurrences differ across the four HDS article categories. Table 5 depicts the total number of *settlements*, the *number of articles*, the *average article length* in words, the *average number of settlements* in an article, and the *number of settlements in 100 words*, across the four article categories. The rightmost column (i.e., *settlements/100 words*) in Table 5 is highlighted in grey and is particularly interesting as it demonstrates how dense spatial information is in articles of the four article categories. The second to the fifth column illustrate the variables used to calculate values in the *settlements/100 words* column: the number of *settlements* divided by the number of articles equals *settlements/article*. *Settlements/article* divided by *length* and multiplied by 100 equals *settlements/100 words*.

Table 5: Article and *settlements* characteristics by article categories.

Article categories	Settlements	Articles	Length	Settlements/ article	Settlements/ 100 words
Biographies	132,904	25,202	128	5.3	4.1
Geographical entities	90,965	5,350	422	17.0	4.0
Thematic contributions	23,020	3,067	625	7.5	1.2
Families	17,814	2,569	183	6.9	3.8
Total	264,703	36,188			
Average			218	7.3	3.4

The density of toponyms (i.e., *settlements/100 words*) is highest for *biographies*, *geographical entities*, and *families* in Table 5, as 4 out of 100 words are toponyms in these categories. The relatively high density of toponyms in the *biographies* and *families* category is not surprising, as these article categories contain many places since they describe which places were important to historically important people and families covered in the HDS. *Geographical entities* describe the history of important places in Switzerland and thus a high density of toponyms was also expected. Toponyms are less dense in the *thematic contributions* articles, in which only 1.2 out of 100 words are toponyms. This is unsurprising, because the *thematic contributions* category covers many themes which are not directly related to space (e.g., articles about *units of measurements*) and the articles are long (i.e., 625 words on average) compared to the other article categories, which results in a low *settlements/100 words* value.

For this project, we only considered the most frequent of the 10,565 unique toponyms of the *settlements* feature type. Based on the discussions with historians in the *focus group meeting*, we aimed at including about 200 toponyms in the *spatialized networks* (see *Subsection 5.3.1*). To locate these toponyms, we inspected a list of 250 toponyms which occur most often in the HDS. From this list of 250 toponyms, we excluded all toponyms which were incorrectly identified as toponyms by the spatial information retrieval algorithm. For example, the term *Ruth* occurs 134 times in the HDS and was identified as a small village in the *Canton of Geneva* in the southwestern part of

Switzerland by the spatial information retrieval algorithm. However, the term *Ruth* refers, in all but one of its 134 occurrences in HDS articles, to the common female first name *Ruth*, instead of the small village of *Ruth* in the *Canton of Geneva*, which we discovered by reading the HDS articles in which *Ruth* occurs. As a result, we excluded *Ruth* from the toponym list. Following this procedure for all 250 toponyms, a set of 203 toponyms remained in the list, and this was considered for the spatialization procedure (see *Subsection 5.2.1*). In Table 6, the number of unique toponyms and the frequency of occurrence across the *SwissNames* object categories and population categories according to *SwissNames* (i.e., today's population) are shown.

Table 6: Object categories and frequency of the 203 most frequent toponyms.

Object category	Population	Unique toponyms	Frequency
Municipalities	> 50,000	10	76,507
	10,000-50,000	54	40,338
	2,000-9,999	105	26,112
	< 2,000	26	4,649
Other settlements	> 2,000	3	1,567
	100-2,000	4	953
	< 100	1	122
Total		203	150,248

The 203 selected toponyms occur 150,248 times in total in the HDS, as shown in Table 6, and thus account for 57% of all toponym occurrences (i.e., 264,703, see Table 5) of the *settlements* feature type. In Table 6, the two *SwissNames* object categories *municipalities* and *other settlements* are listed. All *settlements* which are not a *municipality* according to the definition of the *Swiss Federal Statistical Office*[72] are part of the *other settlements* category. The 10 municipalities with a population of more than 50,000 inhabitants account for 51% (i.e., 76,507) of the 150,248 toponym occurrences in Table 6. The 185 municipalities with less than 50,000 inhabitants cover 47% of all toponyms. *Other settlements* only contribute eight unique toponyms and account for 2% of the overall frequency of occurrence of toponyms in the corpus. The fact that highly populated places are covered more in the HDS corpus than places with a low population is expected, as *people make history* and 27,771 out of 36,188 HDS articles (i.e., 77%) are about historically important people or families (see Table 5). The list of 203 toponyms used for this project, including their frequency and object category, is shown in Appendix A.

The spatial distribution of the 203 selected toponyms is visualized on a map of Switzerland in Figure 36. The 20 most frequent toponyms are labeled, and their locations are shown as orange points. The remaining 183 toponyms are depicted as small grey points. Figure 36 illustrates a high density of frequent toponyms in the densely populated *Mittelland* region in the southwestern, central, northern, and

[72] Register of the municipalities on the website of the *Swiss Federal Statistical Office*: http://www.bfs.admin.ch/bfs/portal/de/index/infothek/nomenklaturen/blank/blank/gem_liste/03.html (accessed July 2016)

northeastern parts of Switzerland. A very high density of toponyms is visible in the region of *Zurich* (i.e., *Zürich*) and particularly along the shores of *Lake Zurich*. This region is the largest metropolitan area in Switzerland, with more than one million inhabitants[73]. Furthermore, a high density of toponyms is visible in the metropolitan areas of *Basel* and *Bern* (i.e., capital of Switzerland), and along *Lake Geneva*, where *Genf* and *Lausanne* are located. Toponym density is lower in the southern region of Switzerland, which is dominated by the Alps. Further to the South, the cities *Sitten* and *Lugano* are shown. *Sitten* is located in the *Rhone valley*, and several toponyms are located east and west of Sitten along the *Rhone valley*. *Lugano* is located south of the Alps in the Italian-speaking region of Switzerland.

The spatial distribution of the most frequent toponyms in Figure 36 thus corresponds with the most populated regions of Switzerland and fulfills our expectations that populated places are covered by the HDS the most, as previously mentioned.

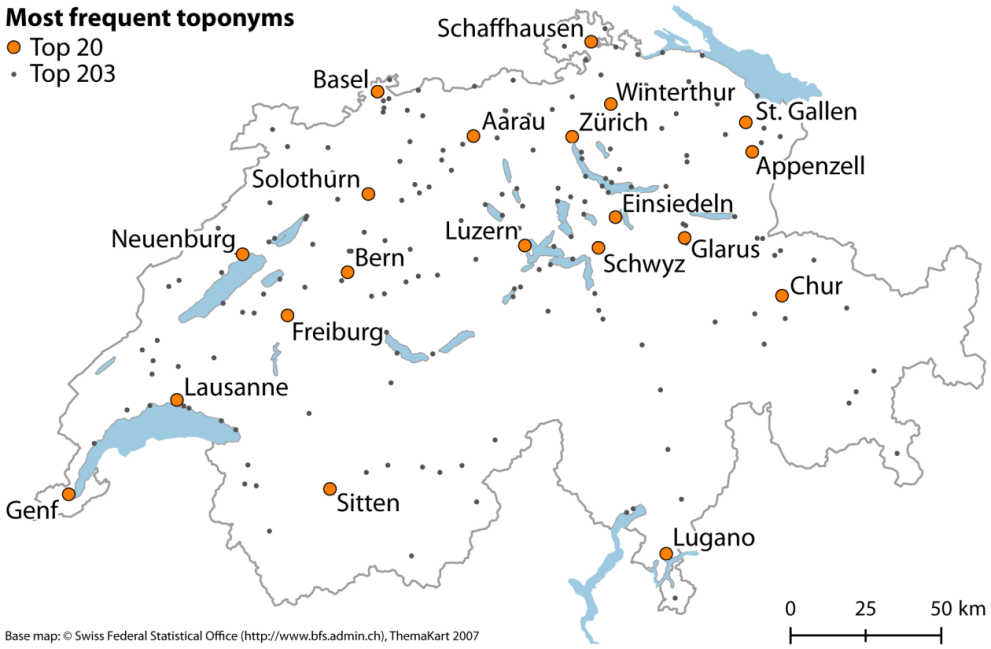


Figure 36: Most frequent toponyms in the HDS.

To summarize, we have illustrated a high density of spatial information in *biographies*, *families*, and *geographical entities* articles of the HDS. We further illustrated that highly populated regions are covered most by the HDS. In the following subsection, we turn to the results of the temporal information retrieval.

5.1.2 Temporal data

We retrieved 499,258 temporal references from the 36,188 HDS articles. In other words, an article contains 13.8 temporal references, on average. We found that 99.5% of all articles contain at least one temporal reference. The high amount of temporal

[73] Metropolitan areas in Switzerland according to the *Swiss Federal Statistical Office*: http://www.bfs.admin.ch/bfs/portal/de/index/regionen/11/geo/analyse_regionen/04.html (accessed July 2016)

references in HDS articles is not surprising as *history happens over time*. For this thesis, we only considered the temporal information from *time* and *date* categories, as we were interested in grouping temporal references to *centuries*, and only references from these two classes can be classified into *centuries*, as mentioned in *Subsection 4.1.2*. These two categories account for 494,077 references and thus cover 99% of all temporal references retrieved from the HDS.

In Table 7, the *number of temporal references* (= *TR*) across the HDS article categories is shown. In addition, the *number of articles*, the *average article length* in words, the *average number of temporal references* in an article (= *TR/article*), and the *number of temporal references in 100 words* (= *TR/100 words*) are depicted across the four article categories. Similarly, as in Table 5, we are most interested in the rightmost column (i.e., *TR/100 words*), which describes the density of temporal information in the four article categories and is highlighted in grey. The second to the fifth column were used to calculate the *TR/100 words* column: the number of temporal references divided by the number of articles equals the *TR/article*. *TR/article* divided by *length* and multiplied by 100 equals *TR/100 words*.

Table 7: Article and temporal references characteristics by article categories.

Article categories	TR	Articles	Length	TR/ article	TR/ 100 words
Biographies	254,505	25,202	128	10.1	7.9
Geographical entities	143,455	5,350	422	26.8	6.3
Thematic contributions	62,042	3,067	625	20.2	3.2
Families	34,075	2,569	183	13.2	7.2
Total	494,077	36,188			
Average			218	13.7	6.3

The density of temporal information (i.e., *TR/100 words*) is highest for the *biographies* article category in Table 7, as 8 out of 100 words are temporal references. In *families* articles, 7 out of 100 words and in *geographical entities* articles 6 out of 100 words are temporal references. The relatively high density of temporal information in *biographies* and *families* articles is unsurprising, as articles in these categories describe the life of historically important people and the history of families; therefore, many dates are listed. Furthermore, articles in these two categories are short (i.e., 128 and 183 words, respectively) which results in a high *TR/100 words* value. The *TR/100 words* value is lowest for *thematic contributions* in Table 7, with 3.2 temporal references in 100 words. This relatively low density compared to the other article categories might be explained by several articles in this category, which contain only few or no temporal references (i.e., articles regarding *units of measurements*), and the articles are long compared to the other article categories, which results in a lower *TR/100 words* value.

In the next step, we aggregated all temporal references to *centuries*, which are illustrated in Figure 37: the y axis indicates the *frequency of occurrence* of temporal references on a base-10 logarithmic scale, while *centuries* are plotted on the x axis. All temporal references referring to the time *before Christ* are summarized in the category *BC*. We

highlighted the *BC* time period in a different color (i.e., in grey) than the 1st to the 21st centuries, as *BC* contains temporal references from several centuries (i.e., all centuries *before Christ*) which are summed together. Therefore, comparing the height of the bar to the bars colored in blue is not meaningful. In addition, we highlighted future centuries (i.e., 22nd and 23rd centuries) in grey in Figure 37, as we manually checked these 23 temporal references in the HDS article texts and discovered that *HeidelTime* (i.e., the tool we used to automatically retrieve the temporal references) retrieved these references by mistake. For example, the notion *2200 BC* in a HDS article was incorrectly annotated as *2200 AD* by *HeidelTime*. This might be due to complex sentence structures, which made it difficult for *HeidelTime* to determine if BC or AD was implied.

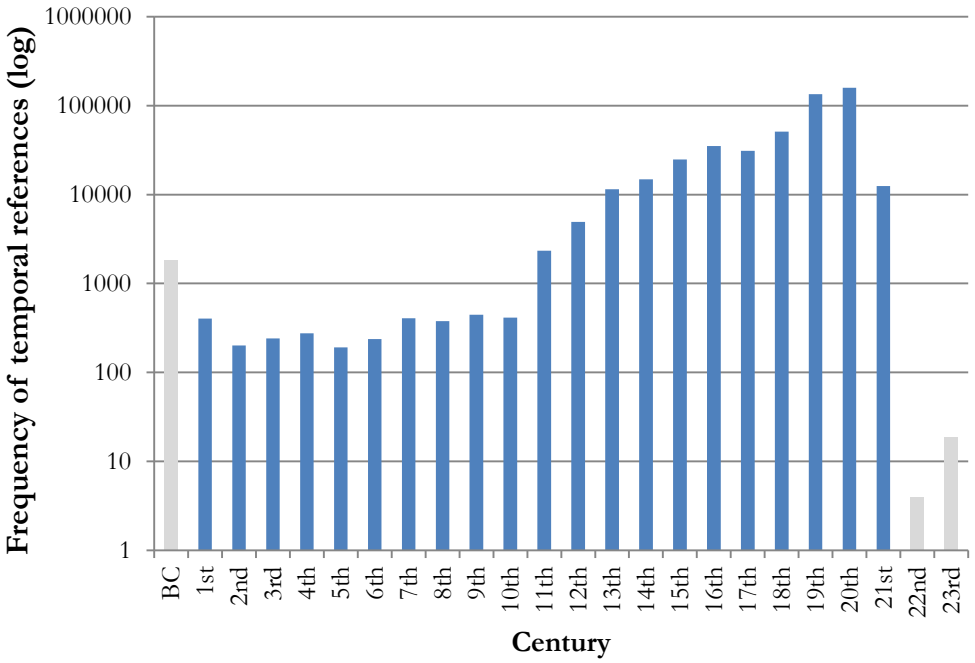


Figure 37: Frequency of temporal references in the HDS by century.

In Figure 37, a general trend is shown: articles regarding the distant past have less temporal information retrieved from the HDS corpus. All centuries between the 1st and the 10th century have a frequency of less than 1,000 temporal references per century. From centuries *before Christ*, 1,846 temporal references were retrieved in total. In contrast, recent time periods are more represented in the HDS, particularly the 18th, 19th, and 20th centuries. In total, 50,988 temporal references were classified as 18th century, 134,329 as 19th century, and 159,027 as 20th century. These three centuries account for 69% of all retrieved temporal references in the HDS. The 19th and the 20th centuries are both overrepresented compared to the initial planning of the HDS, as they planned to assign 20% of all HDS articles to each of these centuries (see *Section 3.1*). However, the 19th century accounts for 27% and the 20th century for 32% of all temporal references instead. For the 21st century, we retrieved 12,462 temporal references.

In this project, we investigate the 18th, 19th, and 20th centuries in detail, as we expected these centuries to be covered most in the HDS corpus (see *Subsection 4.3.2*), which is

confirmed in Figure 37. In Table 8, we illustrate the article and temporal references characteristics by article categories, similar to Table 7. In contrast to Table 7, Table 8 contains only temporal references that are classified as either 18th, 19th, or 20th century. The columns have the same meaning as in Table 7.

Table 8: Article and temporal references characteristics by article categories, limited to the 18th, 19th, and 20th centuries.

Article categories	TR	Articles	Length	TR/ article	TR/ 100 words
Biographies	201,850	25,202	128	8.0	6.2
Geographical entities	83,407	5,350	422	15.6	3.7
Thematic contributions	44,461	3,067	625	14.5	2.3
Families	14,626	2,569	183	5.7	3.1
Total	344,344	36,188			
Average			218	9.5	4.4

By comparing the *TR/100 words* column of Table 8 with that of Table 7, we identify large differences in the *families* and *geographical entities* article categories. For the *families* articles, the *TR/100 words* decreases from 7.2 in Table 7 to 3.1 in Table 8 (i.e., a decrease of 57%). For the *geographical entities* category, the decrease is smaller than in the *families* category, at 42%. Therefore, temporal references regarding the 18th, 19th, and 20th centuries are underrepresented in the *families* and *geographical entities* articles compared to the other article categories, as the decrease in *thematic contributions* is 28% and only 21% in *biographies*. In other words, the 18th, 19th, and 20th centuries are overrepresented in the HDS categories *biographies* and *thematic contributions* when compared to the entire HDS corpus, whereas *geographical entities* and *families* articles are underrepresented. These findings might be explained by the fact that *biographies* and *thematic contributions* particularly focus on people who lived and events which took place in the last few centuries. In contrast, the *geographical entities* articles typically cover as much as is known of the history of a *geographical entity* (e.g., municipality). For example, the article about the municipality of *Zurich* describes its entire history from the first settlements on today's territory of the city of Zurich some thousand years before Christ until today (Behrens et al., 2015). Similarly, articles about *families* typically cover a wide time range, beginning with the first known existence of a family and then reporting on historically important people in the family history. Therefore, the *geographical entities* and *families articles* focus less on the most recent centuries (i.e., 18th, 19th, and 20th centuries) in Swiss history compared to *biographies* and *thematic contributions*.

Based on our analysis, we conclude that the *biographies* article category does not only contribute the largest portion of articles to the HDS, but also has the highest density of spatial and temporal information compared to the other article categories. In contrast, the *thematic contributions* category is characterized by the lowest density of spatial and temporal information. However, we found that the 18th, 19th, and 20th centuries are overrepresented in both the *biographies* and the *thematic contributions* categories when compared to the entire HDS corpus.

Having analyzed the spatial and temporal information in the HDS, we now turn to the thematic information retrieved from the HDS.

5.1.3 Thematic data

In addition to spatial and temporal information, we were also interested in identifying latent topics from the 3,067 HDS articles in the *thematic contributions* category, as we aimed at visualizing the themes that are covered in the HDS corpus and the thematic relationships in a *thematic landscape*, as introduced in *Subsection 4.1.3*. For this reason, we employed the *topic modeling* (TM) technique. Here, we present the output of the TM, since this is used as an input to create the *thematic landscape* (see *Subsection 5.2.2*). Furthermore, we illustrate the thematic diversity of *thematic contributions* articles.

We obtained 30 topics as an output of the TM (see *Section 6.3* for the explanation for why we chose 30 topics), and the most representative terms for each topic as a result of the TM. We translated the three most representative terms for each topic from German to English (see the list in Table 9). A more extensive list of terms is shown in Appendix B in German (i.e., language of the chosen HDS version). Table 9 illustrates that diverse themes are covered by the *thematic contributions* articles. For example, the words of *Topic 2* in Table 9 relate to *politics*, *Topic 4* to *economy*, *Topic 12* to *military*, and *Topic 21* to *agriculture*. As we did not know which themes might be covered by the *thematic contributions* articles due to the lack of thematic metadata in the HDS (see *Section 3.2*), the TM output provided an initial overview of latent topics and the latent structure underlying the HDS corpus.

In Table 10, the probability distributions of four example articles over 11 topics are depicted in order to illustrate the TM output. In each cell of Table 10, the strength of the thematic match between an article and a respective topic is presented as a probability value, which was automatically calculated by the TM software we employed (i.e., MALLET). The higher the probability, the stronger a topic thematically represents an article. In Table 10, only 11 out of the 30 topics are included, as the remaining 19 topics do not thematically explain one or several of the four articles according to the TM output (i.e., probability of 0.0). The first three articles in Table 10 were randomly selected, and the fourth article (i.e., *Matin, Le*) was chosen as it is similar to the third article (i.e., *Journal du Jura*) regarding topic distribution; therefore, it is used as an example to illustrate how the TM output influences the arrangement of thematically similar articles in the *thematic landscape* (as explained in *Subsection 5.2.2*).

The highest probability value for the *Crossair* article in Table 10 is 0.71 for *Topic 15*. *Company*, *found*, and *firm* are listed as most representative terms of *Topic 15* in Table 9, which fits well with the content of the article as it describes that the *Crossair* airline was founded in 1978 and was transformed into *Swiss International Air Lines* in 2002 (Brassel-Moser, 2004). Furthermore, the article describes that bank institutes financially supported the airline in 2001, which might cause the probability value of 0.10 for *Topic 5* which is related to *bank*, *coin*, and *franc*. For five other topics, the probability values for *Crossair* are lower than 0.10 and thus the thematic match is low. The highest probability

value for the article *Militärorganisationen (MO)* (= the organization/structure of the army) in Table 10 is 0.57 for *Topic 12*. The most representative words for *Topic 12* are *army*, *military*, and *federal* which corresponds well with the content of the article, as the article describes the organization of the *Swiss army*, the administration of the *Swiss army*, and military service in general, including military training and education (Senn, 2010). Furthermore, *Militärorganisationen (MO)* has a probability value of 0.32 for *Topic 3*. This topic is related to *federation*, *article*, and *law*. These terms correspond with the content of the article, as the article describes legal aspects of the *Swiss army* since the 19th century (Senn, 2010). For Topics *23* and *28*, the probability values for *Militärorganisationen (MO)* are lower than 0.10 and thus the thematic match is low. Both the articles *Journal du Jura* and *Matin, Le* have very high probability values for *Topic 7*, which contains the words *newspaper*, *appear*, and *journal*. These words are representative of the articles, as they describe daily newspapers in the French-speaking region of Switzerland. For *Topics 3*, *9*, *20*, and *28* the probability values are lower than 0.10 for the *Journal du Jura* and *Matin, Le* articles and thus the thematic match for these topics is low.

Table 9: 30 topics and their most descriptive terms.

Topic	Three most descriptive terms	Topic	Three most descriptive terms
1	Roman, territory, empire	16	trade union, employee, social
2	political, party, found	17	building, example, urban
3	federation, article, law	18	place, Confederation, federal
4	economic, world war, strong	19	parliament, administration, political
5	bank, coin, franc	20	university, school, found
6	woman, child, man	21	agricultural, farmer, agriculture
7	newspaper, appear, journal	22	since, museum, international
8	law, example, court	23	arts, artist, architecture
9	Johann, history, society	24	Italian, German, literature
10	medicine, disease, hospital	25	handcraft, guild, production
11	festival, game, custom	26	building, first, street
12	army, military, federal	27	weight, units of measurement, system
13	international, Federal Council, neutrality	28	first, new, big
14	social, political, rural	29	about, human, culture
15	company, found, firm	30	church, Catholic, Roman

Table 10: Four example articles and their probability distributions over topics.

Article \ Topic	2	3	5	7	9	12	15	20	23	26	28
Crossair	0.02	0.02	0.10		0.03		0.71			0.09	0.03
Militärorganisationen (MO)		0.32				0.57			0.02		0.09
Journal du Jura		0.02		0.92				0.03			0.03
Matin, Le		0.02		0.93	0.01						0.04

The example articles in Table 10 illustrate how well HDS articles are thematically explained by automatically generated topics. Articles which have a similar probability distribution (e.g., *Journal du Jura* and *Matin, Le*) are interpreted by the *self-organizing map* algorithm (see *Subsection 4.2.2*) to be thematically similar and are thus placed close to one another in the *thematic landscape*, which is illustrated in *Subsection 5.2.2*.

To summarize, in this section we first illustrated the spatial and temporal information we retrieved from the HDS articles. We determined that the various article categories differ with regard to the density of spatial and temporal information. *Biographies* is the largest article category with regard to number of articles, and contains a large amount of spatial and temporal information when compared to the other article categories. The thematic information we retrieved from the *thematic contributions* articles illustrate a large diversity of themes covered by the HDS. In the following chapter, we demonstrate how the spatio-temporal and thematic information presented in this section were incorporated in *network visualizations* and the *thematic landscape*.

5.2 Spatialization

We aim at providing an answer to how we can visualize spatio-temporal and thematic structures and interconnections in the HDS in this thesis, and determined *spatializations* to be relevant (as introduced in *Section 4.2*). In order to present spatio-temporal structures and interconnections, *network visualizations* were chosen since we aim to highlight relationships and interconnections between toponyms over time, and to present spatial structures (e.g., hierarchical structures in the toponym network) to interested information seekers in the humanities (i.e., historians in our project). The results of spatializing spatio-temporal information in networks are illustrated in *Subsection 5.2.1*. In order to spatialize thematic structures and interconnections, the *self-organizing map* technique was chosen, which highlights the semantic relatedness of HDS articles and allows interested information seekers in the humanities to rapidly identify regions of similar articles in a map, as introduced in *Subsection 4.2.2*. The results of applying the *self-organizing map* technique to thematic data retrieved from the HDS are presented in *Subsection 5.2.2*.

5.2.1 Network visualization

In this subsection, we present the spatio-temporal networks we created at the country level to display toponym relationships of *Switzerland*, and at the cantonal level to display toponym relationships of the *Canton of Zurich*. At both spatial scales, networks for the 18th, 19th, and 20th centuries are presented. We first illustrate the elements which are depicted in the network visualizations, then describe the structure of the networks qualitatively, and lastly add quantitative measures to confirm the findings regarding network structure.

To compute the relationships of toponyms, we analyzed the co-occurrences of 203 toponyms selected for this project from the HDS articles, including a temporal

weight as well. For example, to calculate the toponym network for the 18th century, we analyzed how often toponyms co-occurred in articles regarding the 18th century. These spatio-temporal relationships were then visualized in *spatialized networks*. Toponyms which possess a strong spatio-temporal relationship are placed close to one another in the network visualizations and are connected with an edge. Details regarding the computation and the visualization of the spatio-temporal relationships are presented in *Subsection 4.2.1*.

The *spatialized networks* of *Switzerland* in the 18th, 19th, and 20th centuries are shown in Figures 38-40, while the *spatialized networks* of the *Canton of Zurich* are visualized in Figure 41. Three elements are explained in the legend of all networks in Figures 38-41: *toponym community*, *centrality*, and *strength of relationship*. In order to delineate *toponym communities*, we applied an algorithm which clusters toponyms that are densely connected and separates toponyms which are weakly connected. Therefore, toponyms which are part of the same *toponym community* possess strong spatio-temporal relationships among one another (see *Subsection 4.2.1* for further details). The *toponym communities* were calculated for each century separately, as we aimed at delineating *toponym communities* which are optimized for each century. Therefore, the number of and composition of *toponym communities* differs for different centuries and are thus not comparable over time. *Centrality*, in Figures 38-41, is calculated by summing the spatio-temporal relationships of a toponym to all other toponyms. The *strength of a relationship* indicates how strong a spatio-temporal relationship between two toponyms is. For the *centrality* and the *strength of relationship* categories in the legends of Figures 38-41, we built four classes per century since we aimed at providing as much detail as possible while still guaranteeing a high level of network readability. The classification is based on the natural breaks algorithm by Jenks (1967). We selected this particular algorithm to show classes of elements with similar values, and to maximize differences between elements of different classes.

In Figures 38, 39, and 40, some toponyms in the network visualizations contain letters (e.g., *a*, *b*) which refer to the two most central toponyms in each *toponym community*. These toponyms are listed in the *toponym community* section of the legend to the right of the network visualizations. The geographic location of these most central toponyms are depicted on a map of Switzerland at the bottom left of Figures 38-40, and on a map of the *Canton of Zurich* at the top of Figure 41. The color of the toponyms in the map corresponds to the colors of the respective toponyms in the networks. At the bottom right of Figures 38-40, and at the bottom of Figure 41, *network characteristics* are listed, which help to quantitatively characterize the network structure. The first two rows indicate the number of nodes and edges for each network. The third row lists the *average shortest path* (ASP). The *shortest path* represents the lowest number of edges between two nodes in a network. For example, the *shortest path* of *Luzern* (c)-*Schwyz* (m) in Figure 38 contains three edges. The ASP is the average shortest path length of all possible pairs of nodes in a network. ASP has been found to be a robust measure of a network's topology, which is used to identify how *linear* or *centralized* a network is (Albert and Barabási, 2002: 49). The more centralized a network is, the lower the ASP. For

example, the 18th century network of the *Canton of Zurich* (see Figure 41) illustrates a maximal centralized network (i.e., all but one node in a network is connected to one center node). In contrast, Figure 38 illustrates the most *linear* network of the networks in Figures 38-41. We decided to include this global measure to be able to quantitatively compare the topology of different networks to one another. In the fourth row, the normalized total *strength* of the networks is displayed. This value equals the sum of all weighted spatio-temporal relationships in a network, and thus describes how strongly toponyms are connected in the networks in Figures 38-41. Therefore, the more that the 203 selected toponyms co-occur in articles about a specific century, the higher the network *strength* is for that century. The values for *strength* are normalized to the spatio-temporal network of Switzerland in the 20th century because this network has the highest *strength* of all networks in Figures 38-41. Normalization was conducted because we aimed at facilitating a comparison of the *strength* of different networks. Below the *strength*, the influence of *biographies*, *thematic contributions*, *geographical entities*, and *families* to the networks is shown. For example, the more the 203 toponyms co-occur in HDS articles of the *biographies* article category, the higher the percentage of *biographies* is in the network characteristics section in Figures 38-41. The spatio-temporal relationships used to calculate the *strength* and influence of the article categories are weighted by the spatial (i.e., *Okapi BM25*) and temporal (i.e., *centuries weights*) criteria demonstrated in Figure 27.

In the following, we first analyze the networks of *Switzerland* in Figures 38-40. After having described the structures and patterns in the network visualizations qualitatively, we underpin our findings with the quantitative measures (i.e., *network characteristics*).

The networks in Figures 38-40 generally appear to be centralized. Therefore, central nodes exist, which are located in the middle of the networks: for example, *Basel* (a) and *Luzern* (c) in Figure 38, *Zürich* (a), *Bern* (g), and *Luzern* (k) in Figure 39, and *Zürich* (c) and *Bern* (k) in Figure 40. Many less central nodes are directly connected to these *network hubs*. Upon comparing the networks in Figures 38-40 with one another, the network in Figure 38 appears most linear and Figure 40 most centralized, since, in Figure 38, a long path from *Bern* (g) at the bottom of the network to *St. Gallen* (j) at the top of the network is illustrated, whereas in Figure 40 such a long path cannot be found, and many nodes are directly connected to the central nodes *Zürich* (c) and *Bern* (k). The network structure in Figure 39 is somewhat in between that of the two network structures in Figures 38 and 40.

Next, we turn to the most central toponyms in the *toponym communities*, which we labeled in the network visualizations and placed on a map of Switzerland at the bottom left of Figures 38-40. For the network in Figure 38, related to toponym relationships in Switzerland in the 18th century, we focus on three findings that are relevant for the general structure of the network: first, *Basel* (a) and *Pruntrut* (b) are located in the middle of the network. For *Basel*, one of the largest cities in Switzerland, the central position is not surprising. In contrast, *Pruntrut* is a city with currently less than 7,000 inhabitants and approximately 2,000 inhabitants in the 18th century (Kohler, 2013); therefore, another argument other than population must be found to explain its central location in the network. The most obvious explanation is that the bishop of the *Prince-Bishopric of*

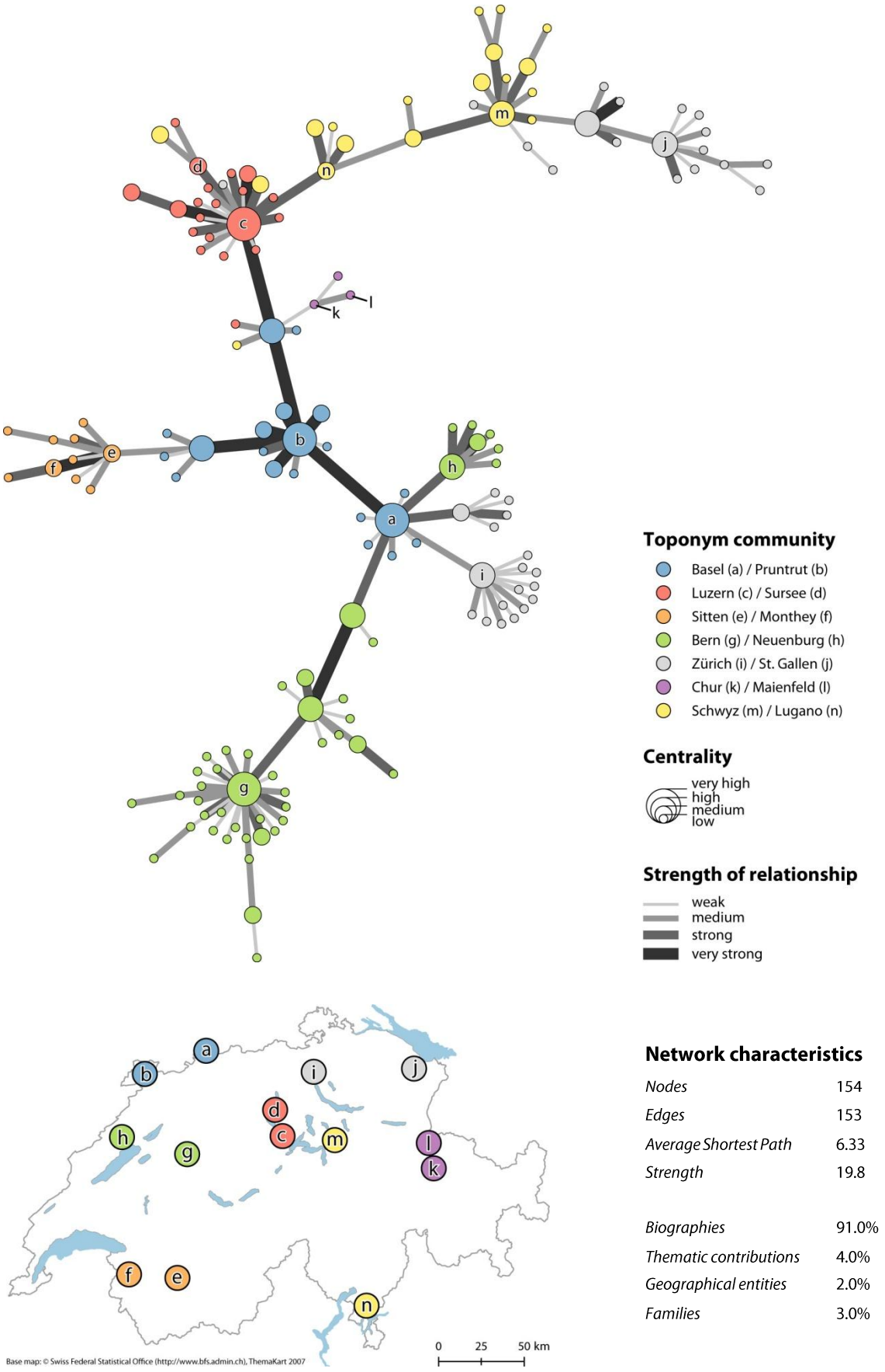
Basel (= *Fürstbistum Basel*), which had feudal authority over large territories in the region of *Basel*, was seated in *Pruntrut* from 1528 until 1792, a fact that is well documented in the HDS (Bandelier et al., 2009). Prior to 1528, the bishop was seated in *Basel*, but due to the reformation in Switzerland in the 16th century, the seat moved to *Pruntrut* (which is located west of *Basel*) (Bandelier et al., 2009). We discovered this by reading HDS articles about the 18th century in which *Pruntrut* occurs. The second finding is that *Schwyz* (m) and *Lugano* (n) are part of the same *toponym community*, and are placed close to one another in the network visualization in Figure 38. This is surprising as there is a large geographical distance between *Schwyz* and *Lugano* as shown in the map of Figure 38 and because *Schwyz* is located in the German-speaking region of Switzerland north of the Alps, while *Lugano* is located in the Italian-speaking region south of the Alps. This can be explained by territory in the Italian-speaking region of Switzerland (= *Ennetbirgische Vogteien*) being owned by the regions north of the Alps at that time (i.e., 1512-1798), which is well documented in the HDS (Hubler, 2004). We detected this by reading articles about the 18th century in which *Schwyz* and *Lugano* co-occur. A third finding we point out is that *Zürich* (i) and *Bern* (g) are located peripherally in the network in Figure 38 compared to Figures 39 and 40. In Figure 38, several nodes that are located in the territory which was owned by the cities of *Zürich* and *Bern*, respectively, at that time (i.e., 18th century), are directly connected to *Zürich* and *Bern*. However, *Zürich* and *Bern* are not central *network hubs* in the 18th century when compared to the networks in the 19th and 20th centuries (i.e., Figures 39 and 40) in which they are connected to toponyms from all regions of Switzerland. This could be explained by the political and economic influence of *Zürich* and *Bern* being lower in the 18th century when compared to later centuries; therefore, regional structures and relationships were more important at that time and are thereby covered in more detail by the HDS.

In the next step, we describe the spatio-temporal toponym relationships of Switzerland in the 19th century (see Figure 39), and highlight two relevant findings regarding the general structure of the network. First, we see that *Zürich* (a) and *Bern* (g) are much more centrally located in this network than in Figure 38. The central position of *Zürich* in the network of Switzerland might be explained by *Zürich* becoming an important industrial and growing financial hot spot for Switzerland in the 19th century (Behrens et al., 2015). Furthermore, *Zürich* became an important hub for traffic and transportation due to the construction of railroads in the region of the city of *Zürich* (Behrens et al., 2015). *Bern* had become the capital of Switzerland in 1848 (i.e., the year when Switzerland became a federal state), which would explain *Bern*'s central position in the network. A second finding from Figure 39 is the connection between *Freiburg* (k) and *Luzern* (l). Both cities are part of the violet *toponym community* and are directly connected in the network visualization, indicating a strong relationship. Both cities are traditionally catholic and both cities have large universities. These factors were identified as the main reasons for this strong relationship, as the cities co-occur often in articles about people who lived in one of these cities and studied in the other, or in *thematic contributions* articles which are related to religion and *Catholicism* (e.g., *Kolpingwerk*).

In the next step, we studied the spatio-temporal toponym relationships in Switzerland in the 20th century (see Figure 40), and highlight two findings. First, Figure 40 illustrates that the positions of *Zürich* (c) and *Bern* (k) in the 20th century are even more central than in the 19th century (i.e., Figure 39). The position of *Zürich* is more central than that of *Bern* (i.e., *Zürich* is directly connected to more nodes than *Bern*). This could be due to the increase of *Zürich*'s importance as economic as well as banking and finance center of Switzerland in the 20th century (Behrens et al., 2015). Furthermore, the *University of Zurich* (i.e., largest university in Switzerland) and the *Swiss Federal Institute of Technology in Zurich* (ETH Zurich) are located in the city of *Zürich*. Both were founded in the 19th century and have become leading international research institutes and thus have established *Zürich* as an important Swiss and international research hot spot (Behrens et al., 2015). The second finding relates to *Freiburg* (i) and *Luzern* (j), remaining part of the same *toponym community* in the 20th century, yet *Freiburg* is directly connected to *Bern*, and *Luzern* to *Zürich* in Figure 40. Therefore, *Freiburg* and *Luzern* remain strongly connected in the 20th century, but the relationships *Bern-Freiburg* and *Zürich-Luzern* are stronger than *Freiburg-Luzern*; therefore, *Freiburg* and *Luzern* are located in different regions of the network visualization in Figure 40.

Until this point, we have qualitatively analyzed the networks. We now turn to the quantitative measures listed in the *network characteristics* sections in Figures 38-40. The number of nodes and edges is lowest for the 18th century and highest for the 19th century. We expected the 18th century to contain the fewest nodes and edges as it represents the century with the fewest temporal references in the HDS when compared to the other two centuries, as demonstrated in *Subsection 5.1.2*. However, the fact that the 19th century contains more nodes and edges than the 20th century is surprising because we retrieved more temporal references from the 20th than from the 19th century (see *Subsection 5.1.2*). The reason for this could be due to the strong dominance of *Zürich* and *Bern* in the 20th century, while toponyms which have only few and relatively weak relationships when compared to *Zürich* and *Bern* might not fulfill the criteria to be incorporated in the toponym relationship network (see *Subsection 4.2.1* for the criteria).

Furthermore, we analyze the ASP values in Figures 38-40. The theoretical maximum for a network containing 198 nodes is 66.3 (i.e., completely linear structure), while the theoretical minimum is 2.0 (i.e., a completely centralized structure). For all networks in Figures 38-40, the theoretical maximum and minimum are similar and thus we do not detail them for other network sizes. We further calculated a randomized network, with 198 nodes, applying the same layout algorithm as for Figures 38-40 (i.e., *pathfinder network scaling*) and obtained an ASP of 11.8. Interpreting the ASP values in Figures 38-40 reveals that all networks are more centralized than the randomized network as the values are between 4.1 for the 20th century and 6.3 for the 18th century. Comparing the three networks, the 18th century network is the most linear because it has the highest ASP, whereas the 20th century network is most centralized as it has the lowest ASP. This confirms our expectations regarding the network structure previously stated.



Network characteristics

Nodes	154
Edges	153
Average Shortest Path	6.33
Strength	19.8
Biographies	91.0%
Thematic contributions	4.0%
Geographical entities	2.0%
Families	3.0%

Figure 38: Spatio-temporal network of Switzerland in the 18th century.

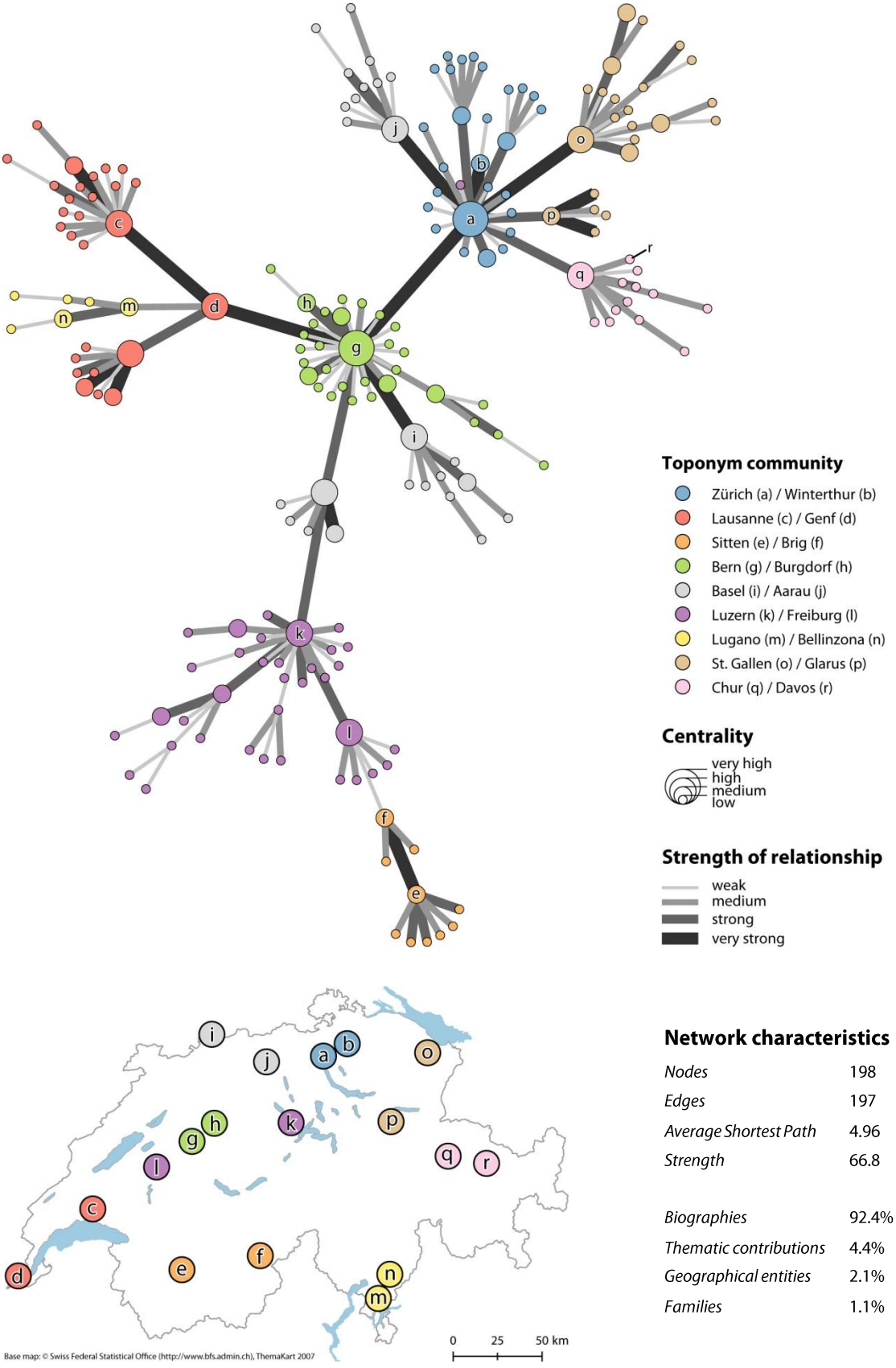


Figure 39: Spatio-temporal network of Switzerland in the 19th century.

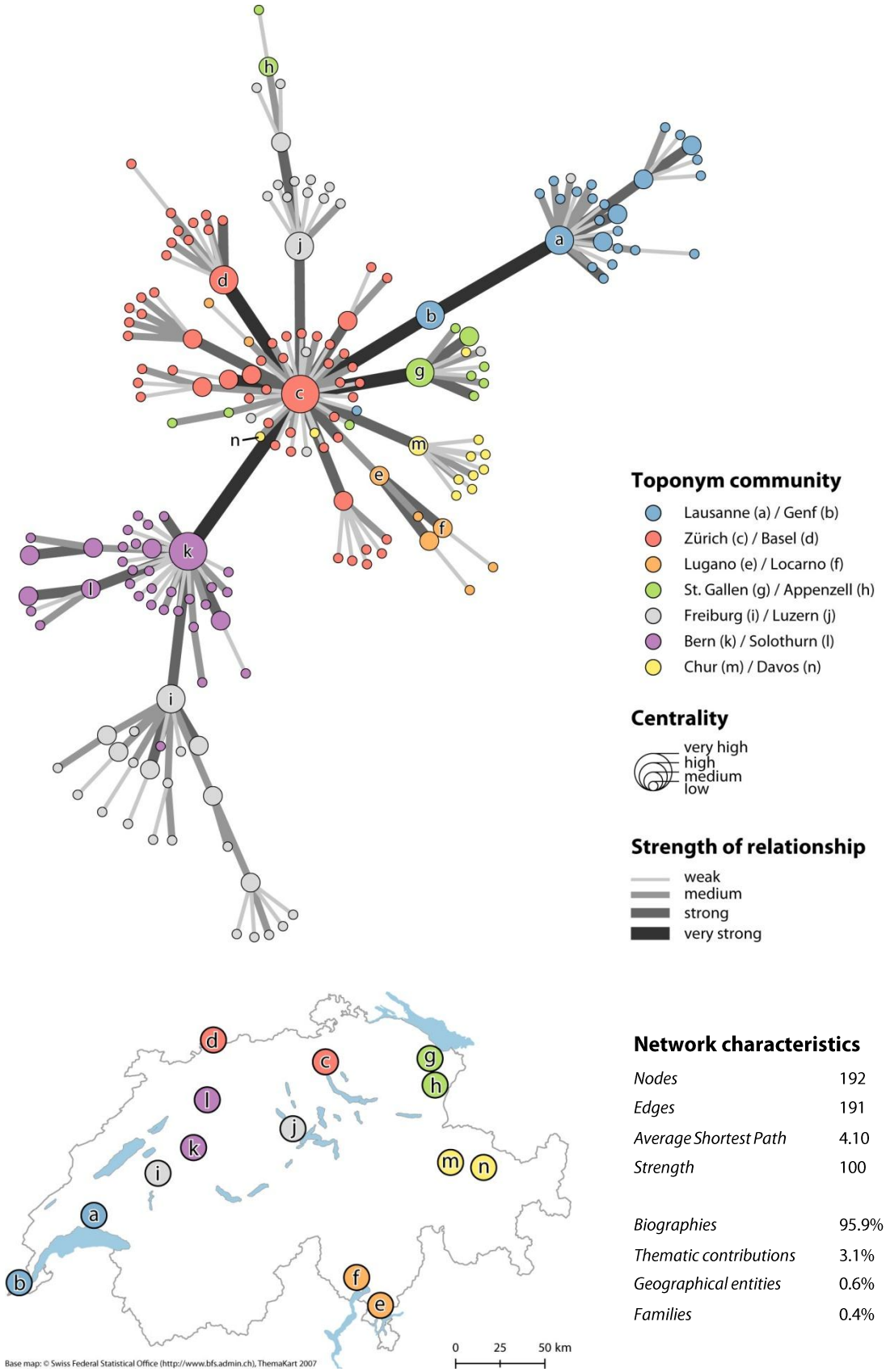
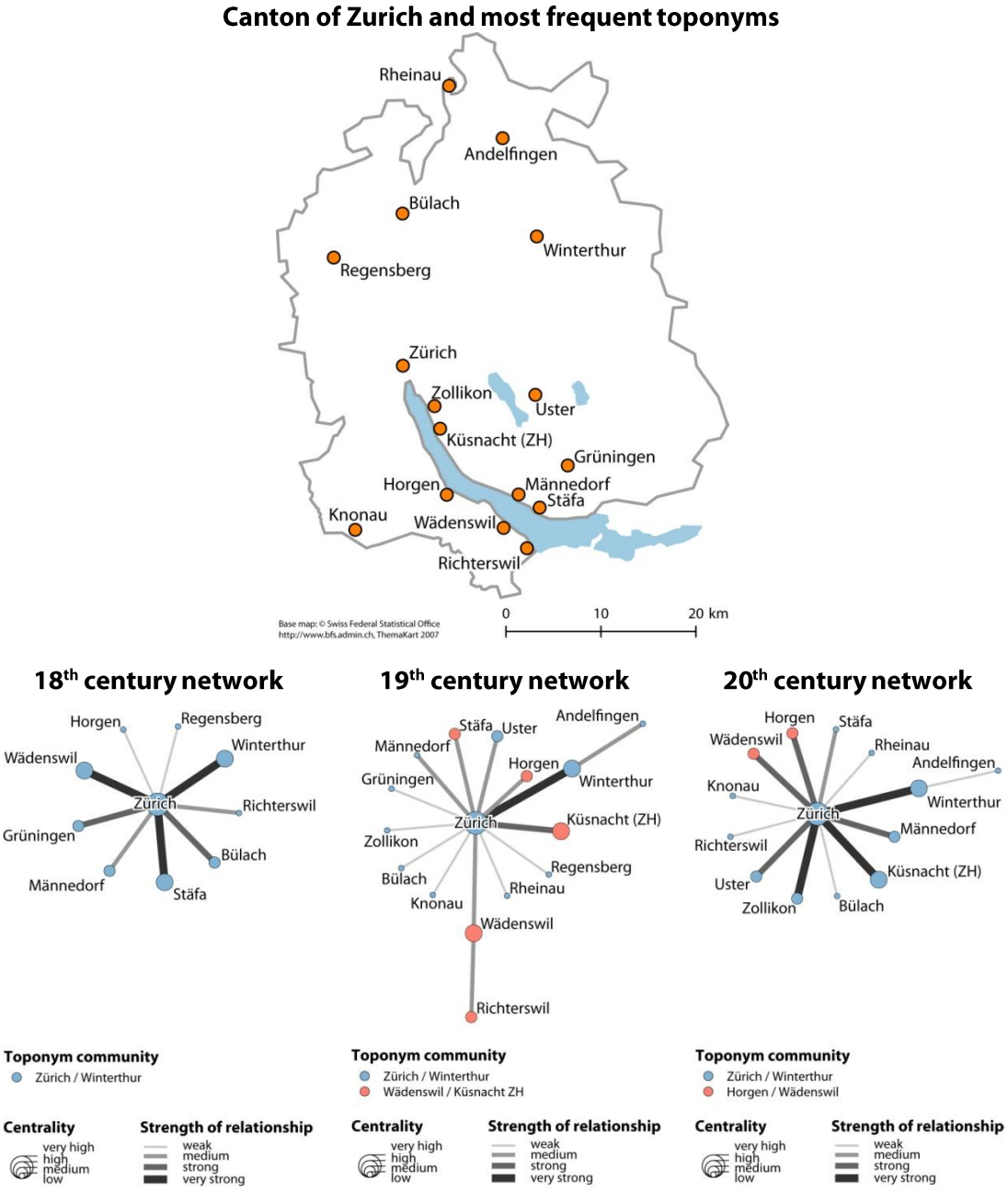


Figure 40: Spatio-temporal network of Switzerland in the 20th century.



Network characteristics

Nodes	10	*Nodes*	16	*Nodes*	14
Edges	9	*Edges*	15	*Edges*	13
Average Shortest Path	1.80	*Average Shortest Path*	2.09	*Average Shortest Path*	1.98
Strength	0.3	*Strength*	2.6	*Strength*	4.3
Biographies	96.2%	*Biographies*	96.4%	*Biographies*	98.5%
Thematic contributions	2.3%	*Thematic contributions*	1.6%	*Thematic contributions*	1.3%
Geographical entities	1.5%	*Geographical entities*	1.4%	*Geographical entities*	0.1%
Families	0%	*Families*	0.6%	*Families*	0.1%

Figure 41: Spatio-temporal networks of the 18th, 19th, and 20th centuries for the *Canton of Zurich*.

Analyzing the network *strength* (i.e., all weighted spatio-temporal relationships in a network) in Figures 38-40 reveals that the 18th century has the lowest *strength* and the 20th century has the highest *strength*. This is what we expected because the most temporal references were retrieved from the 20th century and the fewest from the 18th century (see *Subsection 5.1.2*).

Lastly, we analyze the influence of *biographies*, *thematic contributions*, *geographical entities*, and *families* to the networks in Figures 38-40. It is obvious that *biographies* dominate all networks (i.e., toponyms co-occur most in *biographies*) since the percentage is higher than 90% for all networks and increases over time. Interestingly, although the *geographical entities* article category contains more spatial and temporal information than the *thematic contributions* articles (see *Subsections 5.1.1* and *5.1.2*), *thematic contributions* articles contribute more to the spatio-temporal networks in Figures 38-40. This is primarily due to the fact that *geographical entities* articles cover longer periods of time than a single century, and thus often do not fulfill the minimum criteria of at least 50% of all temporal references in an article needing to be about a century in order to be incorporated in the network of that century (as explained in *Subsection 4.2.1*). The *families* article category contributes the least to the spatio-temporal relationships in Figures 38-40, with the exception of the 18th century. Similar to the *geographical entities* category, *families* articles cover extended time periods and often do not fulfill the 50% criteria for incorporation into the spatio-temporal networks of the 18th, 19th, and 20th centuries; therefore, *families* articles contribute little to the networks in Figures 38-40.

Having illustrated the findings for networks about Switzerland, we now turn specifically to the *Canton of Zurich*. In Figure 41, the networks of the *Canton of Zurich* in the 18th, 19th, and 20th centuries are illustrated. The networks in Figure 41 contain few nodes when compared to the networks about *Switzerland* in Figures 38-40, as only 16 out of the 203 toponyms which were considered in this project are located in the *Canton of Zurich*. As shown in the map at the top of Figure 41, most of these 16 toponyms are located around *Lake Zurich* (i.e., the largest lake in Figure 41). As for the networks about Switzerland, we detail qualitative followed by quantitative findings.

The structure of the 18th century network is maximal centralized, as all but the node of *Zürich* are directly connected to *Zürich*. This pattern was already revealed in Figure 38, wherein *Zürich* is directly connected to all toponyms which were owned by the city of *Zürich* in the 18th century; therefore, the structure is as expected. For example, these relationships are caused by *biographies* about *bailiffs* (i.e., *Landvögte*) in the HDS who were sent by the city of *Zürich* to a region that was owned by *Zürich* at that time. These *bailiffs* were responsible for the administration, judicature, and armed forces in these regions (Hörsch, 2008). We detected this by reading articles in which *Zürich* and the other toponyms connected to *Zürich* in the 18th century network in Figure 38 co-occur.

The structure of the 19th century network is similar to the 18th century network, though *Andelfingen* is connected to *Winterthur*, and *Richterswil* to *Wädenswil*, compared to the other toponyms, which are directly connected to *Zürich*. *Richterswil* is geographically closer to *Wädenswil* than to *Zürich*, and *Andelfingen* is geographically closer to *Winterthur* than to

Zürich, which might represent an initial explanation. Secondly, articles in which *Richterswil* and *Wädenswil*, and *Winterthur* and *Andelfingen*, respectively, occur together are particularly about people who lived in both places during their life or worked in one place (e.g., *Winterthur*) and lived in the other (e.g., *Andelfingen*), which we discovered by reading articles about the 19th century in which the toponyms co-occur. Important factors for that *Zürich* is directly connected to all but two toponyms in the 19th century are the *University of Zurich* and the *ETH Zurich*, both of which were founded in the 19th century, as previously described. Many people grew up in a village or city outside of *Zürich* and then went to *Zürich* to study. In the HDS, articles about historically important people contain both the place where the people grew up and the place where they studied (i.e., toponyms co-occur); therefore, a relationship between both places is detected by the toponym relationship algorithm applied in this project (see *Subsection 4.2.1*). For the 19th century, two *toponym communities* are visualized in Figure 41. The blue community contains more toponyms than the red. The red community contains nodes which are all located along *Lake Zurich*, which indicates strong connections between these toponyms located in the same geographical region, and thus complies with the *first law of geography* which states that *close things are more related than distant things* (Tobler, 1970).

The 20th century network has a very similar structure compared to the 19th century, but a noticeable difference is evident in *Richterswil* being missing in the network, and thus *Andelfingen* is the only toponym in the network (except for *Zürich*) that is not directly connected to *Zürich*. The relationships between *Zürich* and all other toponyms in the 20th century are specifically based on HDS articles including the life stages of people who were described as studying, working, and/or living in *Zürich*. This is not surprising because *Zürich* has become an important economic and research hot spot in the 20th century, as previously discussed.

Regarding the network characteristics section in Figure 41, the same findings can be made as previously stated for the networks about *Switzerland*: the network for the 18th century contains the lowest number of nodes and edges, while the 19th century network contains most nodes and edges. Network *strength* increases from the 18th to the 20th century. The explanation for these patterns, which were provided for Figures 38-40, are also valid for Figure 41 and are thus not repeated here. All ASP values are extremely low, which is due to the very centralized structure of the networks. The contribution of the different article categories is also similar to Figures 38-40, but the dominance of the *biographies* category is even more striking as the *biographies* category contributes more than 96% to all three networks in Figure 41.

To summarize our findings, we state that the toponym networks of Switzerland in the 18th, 19th, and 20th centuries are strongly, and the networks for the *Canton of Zurich* are very strongly centralized. In the 19th and 20th centuries, *Zürich* and *Bern* represent the most central nodes of the networks, which is as we expected because *Zürich* is the economic and financial center, while *Bern* is the capital and political center of Switzerland. Based on the findings of this subsection, we state that networks are useful for highlighting hierarchical structures in spatio-temporal data. We further illustrated that some toponyms which are distant from one another in geographical space are

connected in the spatio-temporal networks (e.g., *Schwyz* and *Lugano* in the 18th century), which is interesting as it contradicts the *first law of geography*: that *everything is related, but near things are more related than distant things* (Tobler, 1970). We further illustrated that *biographies* contribute the most to all networks.

Having illustrated the visualization of spatial and temporal information in *spatialized network* visualizations, we now turn to the visualization of thematic information in a *thematic landscape*.

5.2.2 Thematic landscape

As participants of a *focus group meeting* (i.e., historians) discussed our ideas to implement interactive web interfaces and required that we incorporate thematic in addition to spatio-temporal information access functionalities, we decided to apply the *self-organizing map* technique to highlight the semantic relatedness of the 3,067 *thematic contributions* articles in the HDS (see *Subsection 4.2.2*). This article category covers important themes in Swiss history and contains much thematic information. We further decided to depict two hierarchical levels of information in the *self-organizing map* which should both be incorporated in an interactive web interface: an *overview* and a *detail view*. This complies with Shneiderman's (1996) *visual information-seeking mantra*: "*overview* first, *zoom* and *filter*, then *details-on-demand*". Furthermore, this allows to implement an interface based on two views on the data: a *distant* and a *close reading* view. Both views are detailed here and incorporating them in a web interface is detailed in *Subsection 5.3.2*.

In Figure 42, both an *overview* and a *detail* view of the *self-organizing map* (i.e., *thematic landscape*) are depicted. The *overview* map represents a typical *distant reading* view, according to Moretti (2005), as it allows users to get an impression of the overall thematic structure of the HDS articles, whereas the *detail* map represents a *close reading* view since it presents information at the individual article level. The *detail* map in Figure 42 is an inset map for the area that is framed by a black rectangle in the *overview* map. In this *detail* map, neurons (i.e., hexagons) and articles (i.e., points) are visualized. The arrangement of the neurons and articles is based on the *article-topic matrix*, which was introduced in *Subsection 5.1.3*. The neurons were assigned a probability distribution over the 30 topics based on the output of *topic modeling* (see *Subsection 5.1.3*) and represent the input data (i.e., the *thematic contributions* articles). For example, as there are many *thematic contributions* articles with a high probability for *Topic 7* in Table 9 (see *Subsection 5.1.3*), which is characterized by the words *newspaper, appear*, and *journal*, many neurons were created with a high probability for *Topic 7*. Neurons which share similar topic distributions (e.g., all neurons which have a high probability for *Topic 7*) are placed close to one another in the *thematic landscape*, following the *distance-similarity metaphor* (Fabrikant et al., 2006) which states that similar items should be displayed close together (see *Section 2.2*). Then, the probability distributions of each of the 3,067 *thematic contributions* articles over the 30 topics were compared to the probability distributions of the neurons, and each article was subsequently placed onto the neuron which had the most similar probability distribution. Consequently, an article about a newspaper (e.g., *Journal du Jura* in the *detail* map of Figure 42) was placed onto a neuron which has a high probability for

Topic 7 (i.e., about *newspaper*, *appear*, and *journal*). Therefore, articles which are placed in the same neuron in Figure 42 are very similar regarding their probability distribution over the 30 topics (i.e., their thematic content). Additionally, the larger the distance between articles (i.e., the more neurons between articles) in the *thematic landscape*, the lower the thematic similarity of the articles. Details regarding the *self-organizing map* method we applied are presented in *Subsection 4.2.2*.

We chose various colors in order to differentiate the 28 *themes* of the *thematic contributions* articles in the *detail* map. In order to calculate the 28 *themes*, we took the *article-topic matrix* as an input and clustered articles which have similar probability distributions over the 30 topics (i.e., which are similar in their thematic content), which is detailed in *Subsection 4.2.2*. For example, all articles in the bottom left corner of the *detail* map in Figure 42 were found to be similar in their thematic content and thus share the same color. Borders of regions with differently colored articles (i.e., border between yellow and violet articles in Figure 42) are visualized with a large grey line in the *detail* map. Some randomly selected articles are labeled at the top right corner of the respective article locations with their title in the *detail* map in order to provide an idea of which articles are placed in this region.

In order to create the *overview* map in Figure 42, we assigned a *theme* to each neuron based on the *theme* of the articles which are located in the neuron. Then, the borders between neighboring neurons, which were assigned the same *theme*, were dissolved. The procedure to create the *overview* map is detailed in *Subsection 4.2.2*. Labels were assigned to the *themes* based on a small user study (see *Subsection 4.3.2*). Below each label in the *overview* map, the number of articles (= n) which belong to a *theme* is displayed. In addition, the locations and labels of the four articles which were discussed in *Subsection 5.1.3* (i.e., *Crossair*, *Militärorganisationen (MO)*, *Journal du Jura*, *Matin, Le*) are highlighted in both map views in Figure 42.

In Figure 42, a *theme* might consist of several disconnected regions in the landscape, though we only labeled the largest part. The occurrence of disconnected regions is an artifact of the method we applied to create the *thematic landscape* in this project (i.e., *self-organizing map* and *Blondel community detection algorithm*, as described in *Subsection 4.2.2*). For example, the theme *Society* consists of a labeled core region in the top right corner. A second smaller region of this theme is located in the bottom left corner of the map between the themes *Literature & language* and *Religion*. Articles in both regions are related to *Society*, but in the core region they are thematically more related to the themes in the top right corner of the *thematic landscape* (e.g., *Conflicts & wars*), whereas articles of the smaller region are more related to the themes located in the bottom left corner of the *thematic landscape*. Reitsma and Trubin (2007) empirically tested whether information visualizations containing such disconnected regions are more difficult to read and more complex to interpret for users compared to information visualizations without disconnected regions. However, no statistically significant difference was found between the visualizations; therefore, disconnected regions do not appear to substantially hinder the process of interpreting information visualizations.

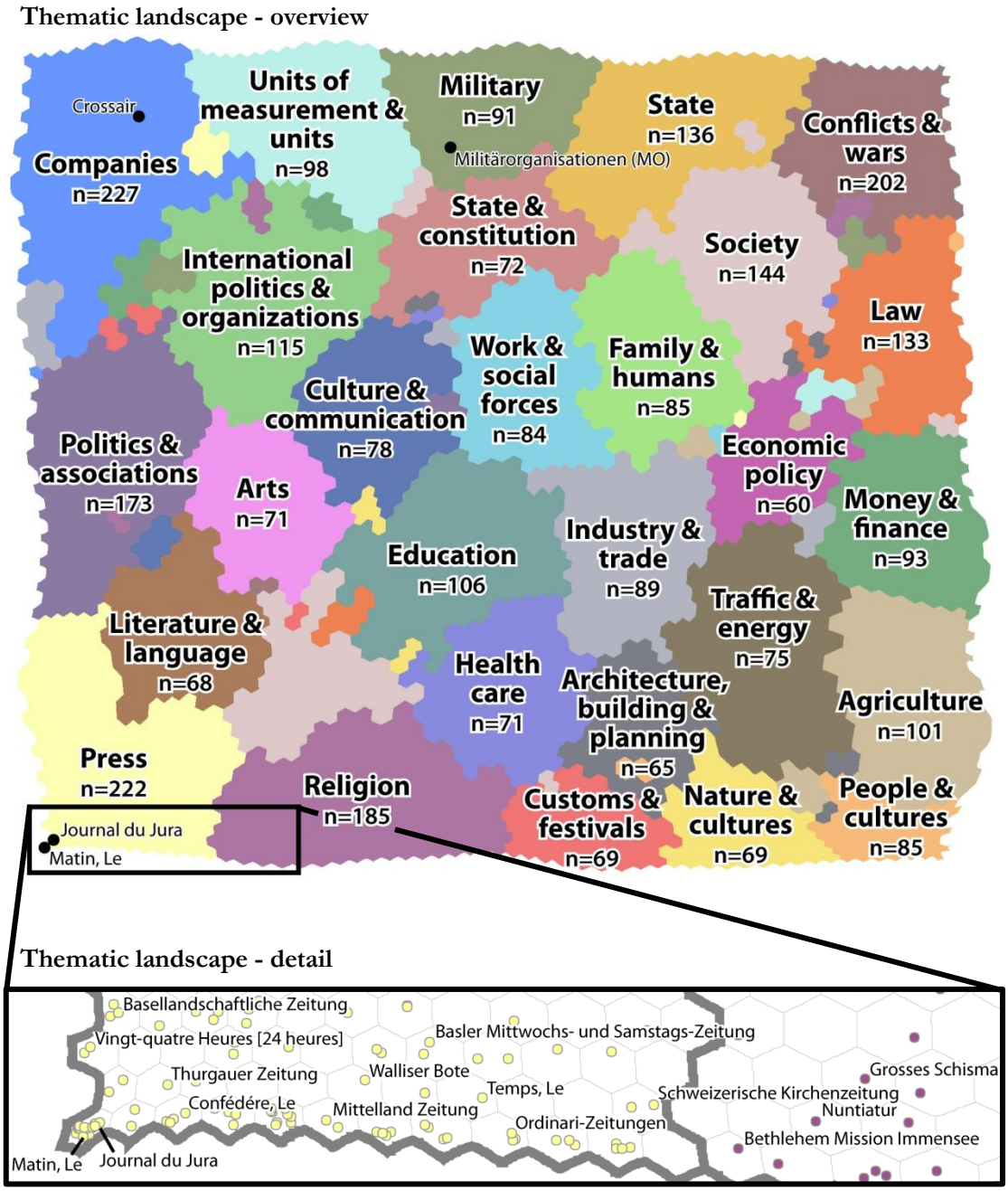


Figure 42: *Overview* and *detail view* of the 3,067 *thematic contributions* articles.

The 28 emerging *themes* in the *overview* map of Figure 42 differ not only in thematic content, but also in size. The smallest thematic cluster contains 60 articles (i.e., *Economic policy*), whereas the largest theme contains 227 articles (i.e., *Companies*). We do not only expect that articles with similar content cluster and form thematic regions (i.e., *themes* in Figure 42), but also that thematic regions of similar content are closer to one another in Figure 42 due to the *self-organizing map* technique we applied. Indeed, in the *overview* map in Figure 42, we identify themes which are similar in content and clustered in the same regions of the map. For example, at the top and in the top right corner of the *overview* map, themes which are related to *state and society* (i.e., the themes *Military*, *State*, *Conflicts & wars*, *State & constitution*, *Society*, *Family & humans*) cluster. At the bottom and in the bottom right corner, we see many *nature and culture* related themes

(i.e., the themes *Customs & festivals*, *Nature & cultures*, *People & cultures*, *Agriculture*). In the bottom left corner, we identify a cluster with the themes *Literature & language* and *Press*, which are both related to *writing and language*. Themes related to *economy and (international) politics* (i.e., the themes *Companies*, *Units of measurement & units*, *International politics & organizations*, and *Politics & associations*) are located in the top left corner. A further *economics* cluster might be identified in the right part of the map, containing the themes *Industry & trade*, *Economic policy*, *Money & finance*, and *Traffic & energy*. As such, our assumption that similar *themes* cluster in the same regions of the map is supported.

Themes which are not clearly attributable to one of these clusters are placed in the center of the map in Figure 42, as they are related to several of the clusters previously mentioned. For example, the theme *Work & social forces* is placed between the *economy and (international) politics* cluster in the top left corner and the *state and society* cluster at the top and in the top right corner. This is reasonable, as the articles of the *Work & social forces* theme (e.g., articles about *labor law*, *strike*, and *social insurances*) are related to *economy*, *politics*, *state* and *society*. Another example is the theme *Education*. *Education* is placed in the middle of the map as it relates to many themes in the *thematic landscape*. Different research fields and the educational system in Switzerland (e.g., articles about universities and schools) are covered in the articles about *Education*. The articles are all related to *education*, but they cover very different other themes as well; therefore, the location in the middle of the map is reasonable.

We expect that articles regarding two neighboring themes in Figure 42 are placed in the border region of these two themes. For example, in the region at the border of the themes *Press* and *Religion* in the bottom left corner of the *overview* map, articles dealing with both themes (i.e., *Press* and *Religion*) should be placed. We consider the *detail* view of this region at the bottom of Figure 42 and see that the article *Schweizerische Kirchenzeitung* is located in this border region, for example. This article is about a religious weekly journal which provides insights into *Catholicism* and the history of *Catholicism* in Switzerland since 1832 (Fink, 2011). Therefore, the location of this article in the border region between the *Press* and *Religion* themes is reasonable.

Furthermore, it is interesting to analyze the locations of the four articles we presented in Table 10 (i.e., *Crossair*, *Militärorganisationen (MO)*, *Journal du Jura*, *Matin, Le*). We reported that *Journal du Jura* and *Matin, Le* have very similar probability distributions over the 30 topics (i.e., they are thematically similar). Therefore, one would expect that they are located in the same region of the map in Figure 42. As Figure 42 shows, both articles are placed in the bottom left of the map and are even assigned to the same neuron, which supports our expectation. Both articles are yellow and thus part of the *Press* theme, which is reasonable considering that they both describe *newspapers*. In contrast, one would expect that articles with very different content (i.e., different probability distributions over topics) such as *Crossair* and *Militärorganisationen (MO)* would not be in close proximity to the *Journal du Jura* and *Matin, Le* articles if the *thematic landscape* followed the *distance-similarity metaphor*. Indeed, *Crossair* and *Militärorganisationen (MO)* are located in different themes, labeled *Companies* and *Military*, respectively, which appears

reasonable as *Crossair* is an airline and the article about *Militärorganisationen (MO)* reports on the organization of the *Swiss army* (see *Subsection 5.1.3*).

To summarize, we demonstrated how we applied the *spatialization framework* to depict the *thematic contributions* articles in a *thematic landscape*, applying a *distant* (i.e., overview) and a *close* (i.e., detail view) reading approach. We specifically illustrated that the arrangement of articles and themes in the *thematic landscape* according to the *distance-similarity metaphor* helps to analyze the structure and the similarities of the articles. Therefore, we found the *self-organizing map* approach useful for the visualization of thematic information and interconnections in a *thematic landscape*.

In this section, we have illustrated the spatialization of spatio-temporal and thematic data in *network visualizations* and in a *thematic landscape*. Now, we turn to the incorporation of these spatialized displays in dynamic and interactive web interfaces.

5.3 Geovisual analytics

In the previous section, we illustrated the results of the spatialization process. In the next step, we incorporated the spatialized displays (i.e., *network visualizations*, *thematic landscape*) in interactive and exploratory web interfaces in order to provide target users (i.e., historians, *digital humanities* interested people) access to the spatio-temporal and thematic data and interconnections we retrieved from the HDS. Interactive web interfaces are particularly useful for this purpose as they allow the depiction of spatio-temporal and thematic data on different hierarchical levels (i.e., *distant* and *close reading*). The design of the interfaces was guided by our target users, as we first discussed initial ideas and evaluated paper mockups of the planned web interfaces with them. The results of these evaluations are detailed in *Subsection 5.3.1*. Based on these results, we implemented prototypes of the web interfaces, which are presented in *Subsection 5.3.2*. We then tested the *utility* and *usability* of the prototype implementations with our target users, reported in *Subsection 5.3.3*. The task lists and mockups are both the input and results of the empirical evaluation process; therefore, they are detailed in this section in order to emphasize the cyclical nature of the design and evaluation process. The participants, the evaluation procedure, and the methods used to analyze the results of the empirical evaluations are presented in *Section 4.3*. Several parts of this section have already been published in Bruggmann and Fabrikant (2016).

5.3.1 Empirical evaluation of the user interface design

We followed a user-centered evaluation and design approach (see *Section 4.3*) in order to develop web interfaces of the *network visualizations* and the *thematic landscape*, which were illustrated in *Section 5.2*. For this reason, we first discussed our initial ideas for the planned interfaces with our target group in a *focus group meeting*. Then, we assessed potential interface design issues by testing hand-drawn paper mockups of the planned web interfaces in a *cognitive walkthrough*. Finally, we presented the hand-drawn mockups that we revised based on the results of the *cognitive walkthrough* to target users in a

think aloud study. The *think aloud study* allowed us to obtain further feedback on the design and planned functionalities of the interfaces. These three evaluation steps are detailed in this subsection and guided the implementation of the prototypes (see *Subsection 5.3.2*).

Focus group research

Mockups and questions

In the first stage of empirical evaluation for the proposed spatialized HDS interfaces, we conducted a *focus group* in order to understand our target users, their needs, and expectations, as discussed in *Subsection 4.3.1*. At this stage, we presented the initial spatialization mockups and sketches of the planned web interface to participants. A sample display containing the 19th century toponym network is illustrated in Figure 43. We further presented networks including the 20th and 21st centuries to participants that are not depicted here, though interested readers are referred to Bruggmann and Fabrikant (2014).

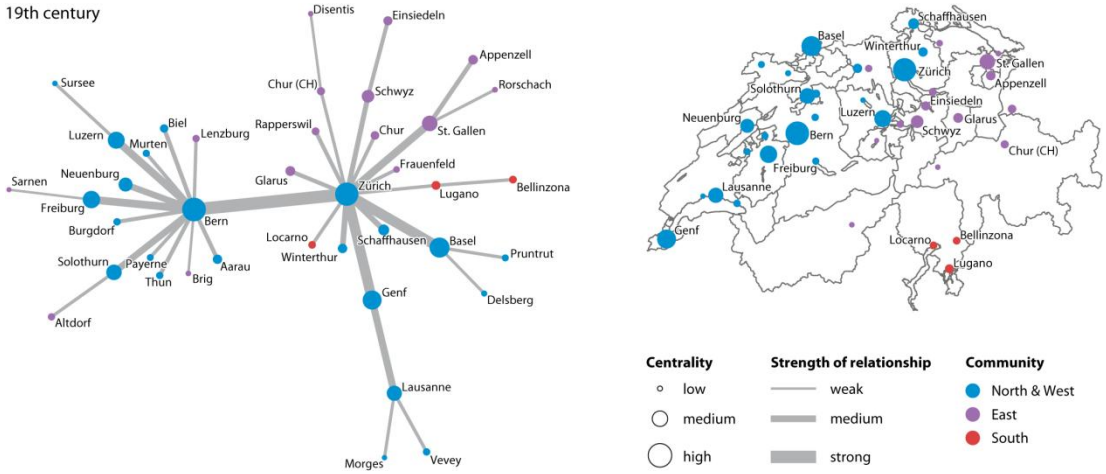


Figure 43: Spatio-temporal network of the 19th century for the *focus group meeting* (modified from Bruggmann and Fabrikant, 2014: 185).

In Figure 43, a spatialization of the 40 most frequent toponyms occurring in HDS articles regarding the 19th century is shown. The spatio-temporal network is based on an earlier version of the spatio-temporal HDS database, as discussed in Bruggmann and Fabrikant (2014). The methods we applied to transform and visualize the spatio-temporal data is equal to the procedure described in *Subsection 4.2.1*.

In Figure 44, a hand-drawn sketch of the initially planned web interface is visualized, which was presented to the *focus group* participants. The reason for using hand-drawn sketches instead of a fully implemented tool was that we aimed to provide target users with the impression that the interface idea is incomplete and non-definite, thereby implying that changes are easily possible at this stage (see *Subsection 4.3.1*). On the left, a menu is shown to help navigate the website. The menu includes options to access further information regarding the entire project, information regarding the data source, and contact details of the author of the website (i.e., the author of this thesis). On the right side, the planned interactive version of a network spatialization is shown.

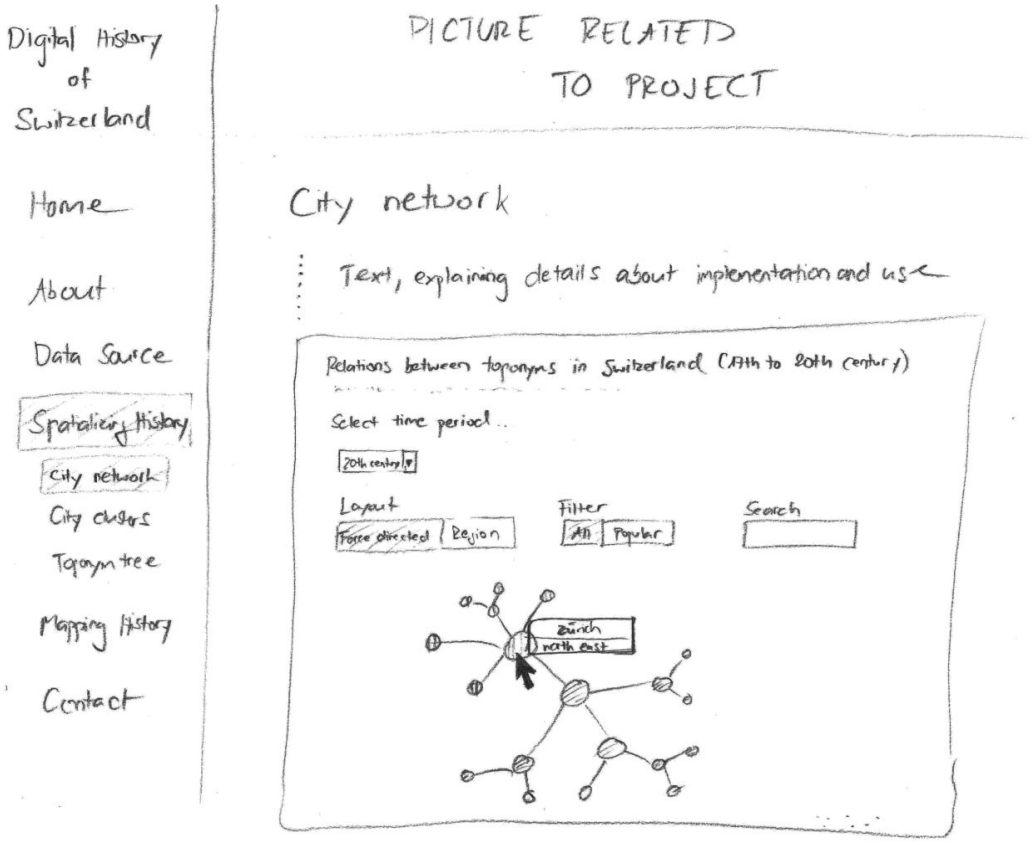


Figure 44: Mockup of the dynamic network visualization for the *focus group meeting*.

The interactive interface in Figure 44 provides functionalities to select a specific time period, to select a network layout, to filter the most frequent toponyms, and a tool to query toponyms in the network. By doing a mouseover in the network, the label of the selected toponym and the respective *toponym community* are displayed.

After providing these visual inputs to *focus group* participants, the author of this thesis initialized a discussion by asking the questions listed in Table 11. The table includes general questions, more detailed questions about the web interface, and other related questions, all asked in order to receive feedback on our initial ideas. Most of the questions are open in order to encourage participants to express their own ideas.

Table 11: *Focus group meeting* questions.

Category	Question
General question	What do you think about the idea to explore Swiss history interactively in space and time?
Spatio-temporal web interface	Would you use a spatio-temporal interface (e.g., dynamic network visualization) to explore Swiss history? And if yes, how often would you use it? Which questions would you like to be answered with such an interface? Which tasks would you like to solve with such an interface?
Other questions	Do you have ideas on how to present spatio-temporal data about history in interactive interfaces?

Results

As a result of this *focus group*, we defined requirements regarding our planned web interfaces and summarized these in four categories: *interactivity*, *transparency*, *knowledge gain*, and *visualization*. Regarding *interactivity*, the participants wished to inspect spatio-temporal data at different spatial and temporal scales (e.g., *zooming* and *time slider* functionalities). Furthermore, they stated a strong interest in querying such a web interface on a higher level of detail: first, a network should include more than 40 toponyms (as shown in Figure 43), and secondly, not only the strongest spatio-temporal relationship of a toponym, but also the second and third strongest relationships should be displayed in an interactive version of the *spatialized network* visualization. In addition, they stated the importance of thematic information in history, and wished to incorporate this in an interactive web interface in order to provide access to the HDS from a thematic point of view. The need for *transparency* was another requirement of the *focus group*. They expressed the need for information regarding the raw data as well as detailed information regarding the computation and visualization of the spatio-temporal and thematic data. Furthermore, they asked for access to source information (i.e., the original HDS articles) while interacting with the interactive web interface in order to judge the validity of the relationships presented in the interface (i.e., coupled *distant* and *close reading* functionalities). As a third point, participants highlighted an interest in *gaining new knowledge*. They referred to the *spatialized network* example and mentioned that the value of such an interactive web interface is increased if unexpected relationships in the data can be uncovered so that new insights from the database and resulting hypotheses might be generated. The fourth point regarding *visualization* refers to participants who asked for specific visualization techniques to be included in a spatialized web interface in order to support spatio-temporal and thematic query tasks. They directly referred to the planned interface in Figure 44, and suggested to complement the network interface with a map.

We incorporated these findings in a six-items-based task list, shown in Table 12. The left column contains the tasks, and the right column contains design implications for the web interfaces. Tasks 1 to 4 refer to the planned interactive spatio-temporal network visualization. We incorporated the *interactivity* requirements mentioned by *focus group* participants by including different spatial and temporal scales for Tasks 1 to 4, and by making a set of the strongest spatio-temporal relationships interactively available. The elicited interface requirement to incorporate thematic information is also reflected by Task 3. Therefore, we create thematic information about the article categories that are also interactively available in the network visualizations. With Tasks 5 and 6, we aim to also include *self-organizing maps* displaying thematic information as an additional component to an interactive multi-view display. The choice of *self-organizing maps* is described in *Subsection 4.2.2*. The requirement to involve more than 40 toponyms in the *spatialized network* visualizations is not reflected in Table 12. However, we considered this point by incorporating a higher number of toponyms (i.e., 203 in total) during the implementation stage of the web interface design phase (see *Subsection 5.3.2*).

Table 12: Task list and design implications (Bruggmann and Fabrikant, 2016: 11).

Task	Design implication
1) Compare the strength of two toponym relationships at a certain spatial/temporal scale	Network visualization as shown in Figure 43 and 44
2) Identify the strongest spatial relationships of a toponym at a certain spatial/temporal scale	Network visualization with an option to show strongest relationships of a toponym
3) Compare the strength of two toponym relationships regarding a specific article category at a certain spatial/temporal scale	Network visualization with an option to analyze toponym relationships according to article categories
4) Compare the community membership of a toponym in two different centuries	Depictions of different temporal network states next to one another
5) Identify articles about a specific topic and thematically similar articles about a specific topic	Visualization of articles in a self-organizing map
6) Identify toponyms that are most relevant to a specific topic	Visualizations of toponyms in a self-organizing map

The *focus group* participants' requirement for *transparency* did not affect the task list in Table 12. We considered this in further steps of the design process of our web interfaces, as we planned to incorporate detailed information about the project, the raw data, and the computation and visualization of the spatio-temporal and thematic data on a separate page of our website. Furthermore, we planned to incorporate source information in the *self-organizing maps interface* (i.e., the *thematic landscape*) as users may access the original articles on the HDS while interacting with the web interface. The *visualization* requirement is also not reflected in the task list in Table 12. We considered this in the further steps of the design process of the *spatialized network interface* by incorporating a map in addition to the network visualization. The last requirement of the *focus group* was that the web interfaces should support the *gaining of new knowledge*. We illustrate how we incorporated this requirement in the prototypes of the web interfaces in *Subsection 5.3.2*.

Summary

We presented the initial spatialization results and sketches of the planned web interface to the *focus group*. As a result of the *focus group* discussion, we decided to implement two web interfaces: an interactive *spatialized network interface* and an interactive *self-organizing map* (i.e., *thematic landscape*). Both interfaces are optimized regarding the *interactivity*, *transparency*, *knowledge gain*, and *visualization* requirements by *focus group* participants. The interfaces are planned to be presented on a website with an additional page that contains further information about the project, the data, and the computation and visualization of the data and relationships.

Cognitive walkthrough

Tasks and mockups

In the next step of the user-centered design and evaluation process, we applied the *cognitive walkthrough* method. The *cognitive walkthrough* is an evaluation method without users and aims at removing as many problems of the planned interface design as possible, prior to target users becoming involved in the evaluation process. For this reason, the designer of the interface (i.e., the author of this thesis) sketched the planned interfaces and tried to decide if target users might interact with the interfaces as expected or not. This procedure is detailed in *Subsection 4.3.1*.

We first had to define representative and specific tasks which can be performed with the planned interfaces, which is necessary to conduct a *cognitive walkthrough* as described here. For this reason, we translated the generic six tasks on the task list we elaborated as a result of the *focus group* (see Table 12) into specific tasks which are presented in Table 13.

Table 13: Task list for the *cognitive walkthrough*.

	Task
1)	Navigate to the *spatialized network interface* and compare the toponym relationship *Bern-Solothurn* with *Bern-Basel* at the spatial scale *Switzerland* and the temporal scale *18th century*. Name the stronger of the two relationships.
2)	Navigate to the *spatialized network interface* and name the three most important toponym relationships of *Basel* at the spatial scale *Switzerland* and the temporal scale *19th century*.
3)	Navigate to the *spatialized network interface* and compare the toponym relationship *Winterthur-Zürich* and *Zürich-Uster* at the spatial scale *Canton of Zurich* and the temporal scale *19th century*. Name the relationship which is stronger regarding the *thematic contributions* articles.
4)	Navigate to the *spatialized network interface* and compare the *toponym community* of *Genf* in the *18th* and the *19th centuries* in the small multiples view. Name the potential difference or state that no difference was found.
5)	Navigate to the *thematic landscape* and find the theme *economy*. Select an article of the *banking and insurances* sub-theme and name the title of the selected article as well as three thematically similar articles.
6)	Navigate to the *thematic landscape* and name the three toponyms which fit best into the theme *economy* in the toponym view. In addition, find the three topics which are most relevant for the toponym *Zürich*.

For prototype implementation, we only considered Tasks 1, 2, 3, and 5 in Table 13. Tasks 4 and 6 were not considered based on the feedback we received in the *target group-based think aloud study I* and are thus colored in grey in Table 13. The reasons for excluding these two tasks are detailed in the following section about the *target group-based think aloud study I*.

For each task in Table 13, we defined a correct series of actions (e.g., click on a button, do a mouseover) before performing the *cognitive walkthrough*. The author of this thesis then drew paper mockups which illustrate the state of the web interface before and after executing an action. We developed a *success* or a *failure story* for each step in the *cognitive walkthrough*, as introduced in *Subsection 4.3.1*. To illustrate this process, we chose

Tasks 2 and 5 as they are typical and representative for all tasks regarding interface issues that we uncovered.

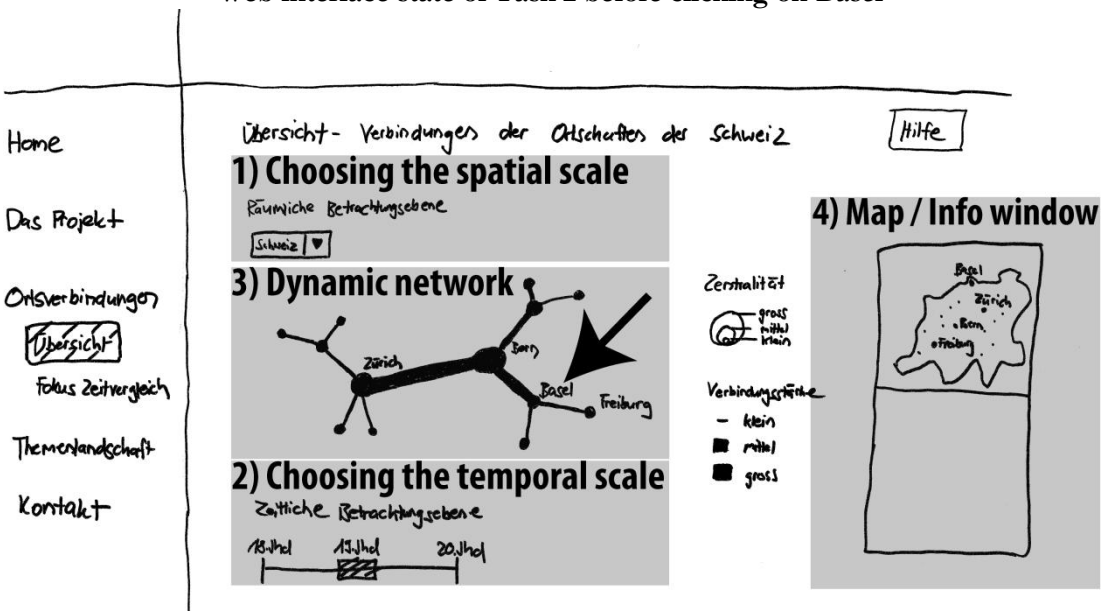
Task 2 consists of different actions, as demonstrated in Table 13. First, the *spatialized network interface* has to be selected and then the strongest relationships of *Basel* at the spatial scale *Switzerland* in the *19th century* must be named. Assuming that the *spatialized network interface* as well as the correct spatial and temporal scale were already selected (numbers 1 and 2 in Figure 45) or are selected by default, the state of the interface looks like in the mockup at the top in Figure 45. Since we are dealing with Swiss history and worked with the German HDS version, we chose German as the interface language.

The appropriate action in Figure 45 is to click on the node of the toponym *Basel* (number 3 in Figure 45, the arrow is pointing to *Basel*) in the network. This user interaction would change the content of the map and the info window (number 4 in Figure 45), and the interface state would look as shown in the mockup at the bottom of Figure 45 (see number 5). As a result, the five strongest relationships of *Basel* are visualized as edges between the toponyms in the network visualization and in the map. The strengths of these five relationships are also listed in the info window below the map. As Task 2 in Table 13 requires only identifying the three strongest relationships of *Basel*, the expected action would be to name the three relationships with the highest strength displayed in the information window (i.e., *Basel-Bern*, *Basel-Zürich*, *Basel-Genf*).

Task 5 consists of different actions, as shown in Table 13. First, the user navigates to the *thematic landscape*. Then, the user selects an article of the *banking and insurances* sub-theme (= *Banken und Versicherungswesen*) within the *economy* theme (= *Wirtschaft*) and names the title of this article as well as the title of three thematically similar articles. Once the user selects the *economy* theme in the *thematic landscape*, the web interface would look like the mockup at the top in Figure 46. On the left, the *thematic landscape* is depicted (number 1 in Figure 46), and zoomed in to the *economy* theme. Above the legend, an inset map illustrates the position of the *economy* theme in the *thematic landscape* (number 2 in Figure 46). The articles are depicted as black points in the thematic landscape. The sub-themes are visualized as regions with different textures and the meaning of the textures is explained in the legend to the right of the *thematic landscape*. The first entry in the legend describes the *banking and insurances* sub-theme.

The appropriate action in Figure 46 is to click on a randomly selected article in the sub-theme *banking and insurances*. A pop-up window for the clicked article then appears. Assuming that the user would choose the *Credit Suisse Group* article, the pop-up window of the *Credit Suisse Group* article would appear (number 3 in Figure 46). As a result, the article title and the article titles of the three most thematically similar articles are displayed, and hyperlinks to the articles in the HDS are listed. As Task 5 in Table 13 requires the user to name the article title of a randomly selected article and the article titles of the three most similar articles displayed in the pop-up window, the correct action for this task would be to name all article titles displayed in the pop-up window in Figure 46 (i.e., *Credit Suisse Group*, *Finanzplatz*, *Banken*, *Wirtschaftspolitik*).

Web interface state of Task 2 before clicking on *Basel*



Web interface state of Task 2 after clicking on *Basel*

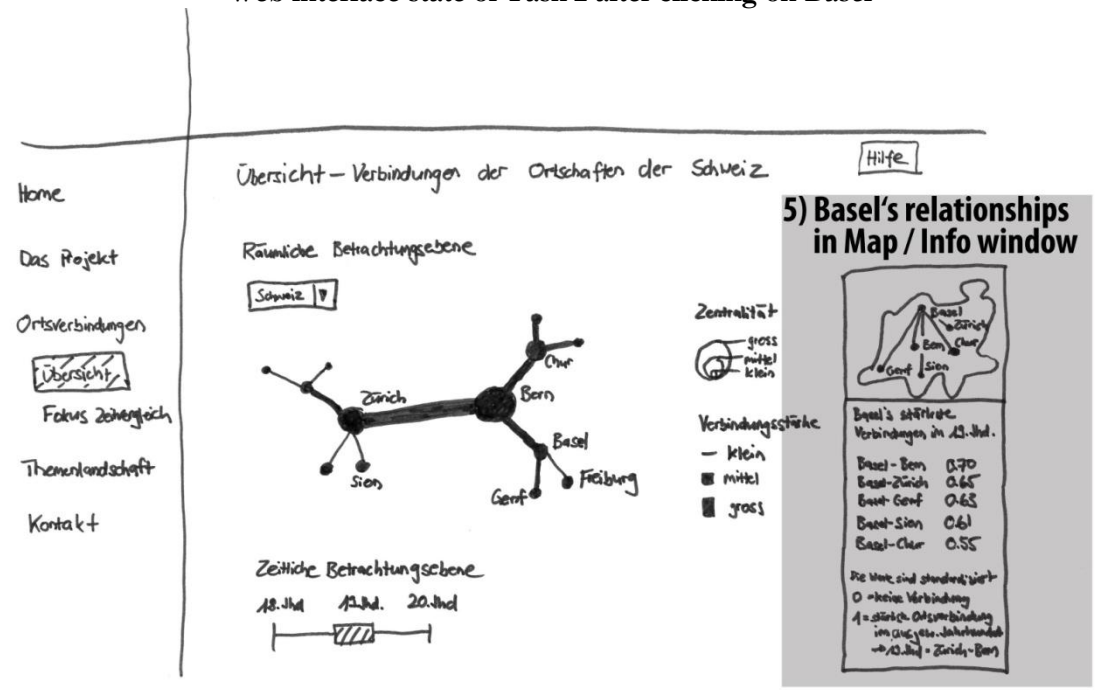
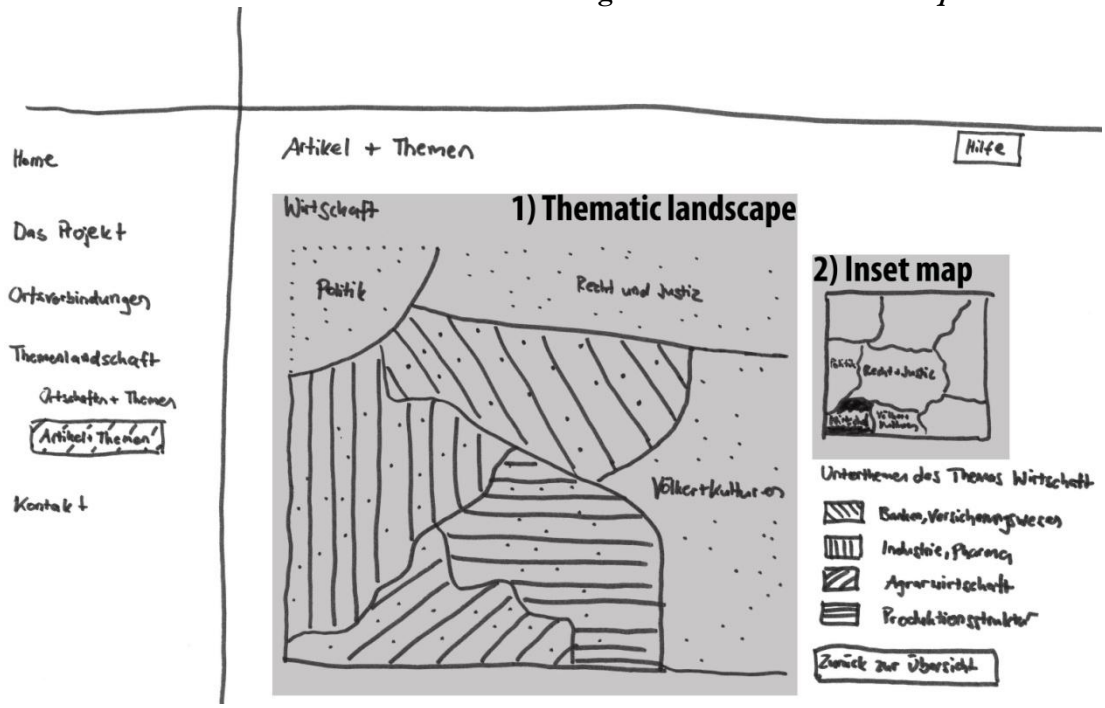


Figure 45: Interface states before and after clicking on the toponym *Basel* in the network visualization (modified from Bruggmann and Fabrikant, 2016: 11).

Web interface state of Task 5 before clicking on the *Credit Suisse Group* article



Web interface state of Task 5 after clicking on the *Credit Suisse Group* article

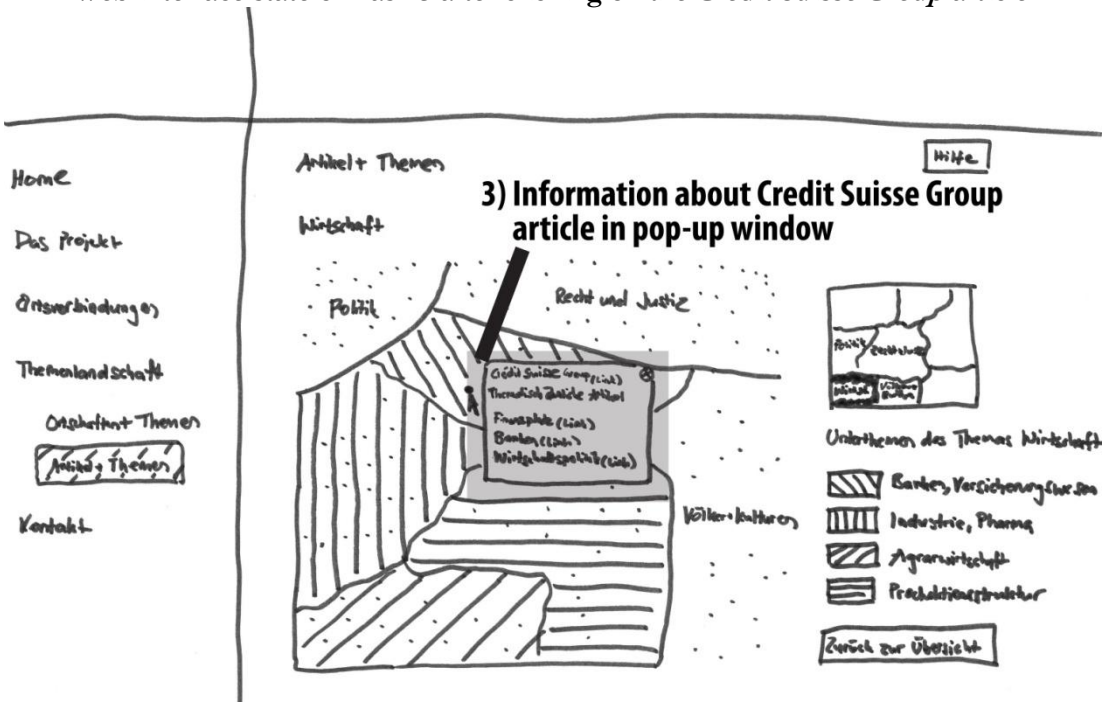


Figure 46: Interface states before and after clicking on the article *Credit Suisse Group* in the *thematic landscape*.

Results

We defined *success* or *failure stories* for all tasks based on predictions of what users might do with the interfaces, as described in *Subsection 4.3.1*. For Task 2, we defined the story presented in Table 14.

Table 14: Action and story for Task 2 (Bruggmann and Fabrikant, 2016: 12).

Action	Outcome	Story
The user clicks on *Basel* in the network	Failure	The user attempts to do a mouseover instead of clicking on the node for the toponym *Basel* in the network visualization. The mouseover does not provide any user feedback. The user attempts to click on the dot at the location of the toponym *Basel* in the map instead of clicking on the node of the toponym *Basel* in the network visualization. The user does not receive any feedback from the system.

As Table 14 depicts, we defined a *failure story* for Task 2 based on two issues: first, we expect that users would rather attempt to do a mouseover than to click on *Basel* in order to see the strongest toponym relationships of *Basel*. Second, a user might expect that it is possible to select the city of *Basel* in the map instead of clicking on *Basel* in the network visualization.

For Task 5, we defined the story presented in Table 15. We identified one issue similar to the issue identified for Task 2 in Table 14. We predict that a user would attempt to use mouseovers in the *thematic landscape* to view article titles. As we only defined the mouse click as the appropriate action and mouseover was not planned to be implemented, the user does not receive any feedback from the system while doing the mouseover.

Table 15: Action and story for Task 5.

Action	Outcome	Story
The user clicks on an article in the sub-theme *banking and insurances*	Failure	The user attempts to do a mouseover in order to see the article titles instead of clicking on an article in the map. The mouseover does not provide any user feedback.

These *failure stories* for Tasks 2 and 5 are representative examples for all tasks that we considered. We found several such interaction issues (i.e., clicking vs. mouseover) potentially leading to failures. Furthermore, we reformulated the task descriptions slightly in order to have clear tasks for our target users in the *target group-based think aloud study I* (see next section). For example, instead of "name the stronger of the two relationships" in the task description of Task 1, we reformulated this part to "as soon as you have named the stronger of the two relationships, you have completed this task". We expected this clarification to be necessary for target users to know exactly when they have accomplished a task.

The results of the *cognitive walkthrough* were used to revise anticipated user *interaction sequences*. The *paper mockups* and *task descriptions* were only slightly adapted.

Summary

We designed paper mockups of the planned web interfaces and defined anticipated user interaction sequences with the interfaces for six tasks. We developed a *success* or a *failure story* for each task based on anticipated user behavior with the interface. We addressed potential sources of user confusion (i.e., *failure stories*) by either changing the anticipated action sequence, the paper mockups, or the task descriptions. We discovered issues mostly related to direct manipulation elements (i.e., clicking vs. mouseover).

Target group-based think aloud study I

Tasks and mockups

In the next step of the user-centered design and evaluation process, we applied the *think aloud method*. In a *think aloud study*, participants are asked to solve tasks and comment on their thoughts while performing tasks in order to identify potential interface design problems. We used the tasks elaborated in the course of the *cognitive walkthrough* study previously detailed and presented paper mockups of the interfaces to target users, depending on the interactions of the participants with the interface. The procedure is reported in detail in *Subsection 4.3.1*.

The task list in Table 13 did not need to be substantially revised as a result of the *cognitive walkthrough*, and thus no changes to the content of the tasks were made. Therefore, we refer to the task list in Table 13 for the *target group-based think aloud study I*.

Tasks 2 and 5 again serve as representative examples for reporting on the type of results obtained from the *think aloud studies*. We again used the paper mockups, which were illustrated in Figures 45 and 46, because no adaptations to the mockups were necessary according to the results of the *cognitive walkthrough*. Based on the *cognitive walkthrough*, a mouseover is the anticipated action to display the five strongest toponym relationships of *Basel* in the network visualization and on the map for Task 2. A user can either do a mouseover on the node of the toponym *Basel* in the network visualization or on the dot at the location of the toponym *Basel* in the map. For Task 5, the action sequence was also adapted and the article titles are displayed if users do a mouseover on the articles in the *thematic landscape*.

Results

Users were asked to solve a list of six tasks (see Table 13) using paper mockups. We observed users while they solved the tasks and compared their actions with the anticipated interaction sequence we defined as a result of the *cognitive walkthrough*.

Table 16 illustrates issues that arose in the *think aloud sessions* for Task 2. We ranked the issues by *importance* and *difficulty* and formulated ideas on how to solve the issues in further versions of the interfaces.

Table 16: Issues, *importance/ difficulty* rating, and ideas to improve the interface for Task 2 (Bruggmann and Fabrikant, 2016: 13).

Issue	Importance/ Difficulty	Fix?
Users do not realize that they chose the wrong spatial or temporal scale.	High importance, medium difficulty	The interaction elements regarding the spatial and temporal scale will be positioned above the network to be more clearly visible.
The network visualization has no zooming function to access multiple spatial hierarchies of the network.	Low importance, high difficulty	No, not of immediate importance, but probably in a further release.

The first issue in Table 16 refers to users who did not realize that they chose the wrong spatial or temporal scale. Three out of five participants in the *think aloud study* had this issue for Task 2, and also for Tasks 1, 3, and 4. We therefore rated this issue as highly important to fix. In the debriefing sessions, one participant suggested placing both interaction elements to choose the spatial and the temporal scale above the network visualization to be more salient. Another issue we identified was that a participant expected a zooming functionality in order to access different spatial scales in the network visualization. We rated this as being of low *importance* as it was mentioned by one participant only. We further rated this issue as highly difficult to implement for interactive network visualizations. Therefore, we stated that we might consider this in a future release.

For Task 5, we identified one issue listed in Table 17. Participants thought that they first needed to click on the *banking and insurances* region in the *thematic landscape*, to be able to zoom in to that region. However, no user feedback is provided for this action by the system. We decided to fix this issue, as we ranked it as highly important. One option to address this issue and reduce user confusion is to implement the typical web zoom functionality available with most common web mapping tools such as *Google Maps*[74].

Table 17: Issues, *importance/ difficulty* rating, and ideas to improve the interface for Task 5.

Issue	Importance/ Difficulty	Fix?
Users cannot click on the sub-theme labeled *banking and insurances* to zoom in to that region.	High importance, medium difficulty	We will implement a typical web mapping zoom functionality.

Similar issues resulted from the other four tasks users were asked to perform. In particular, interface interaction elements were not used by the participants as anticipated in the action sequence. Moreover, participants suggested additional functionality or adaptations in the interface. For example, one user suggested labeling toponyms dynamically in the network visualization, depending on mouseover. Another suggestion

[74] Google Maps: https://www.google.com/maps (accessed August 2016)

was to include pop-up windows in the *spatialized network* to further explain visualization details and content (e.g., *centrality*, *strength of relationship*) on demand. For Task 3, one participant also suggested to display articles in which toponyms co-occur within the information window instead of depicting the contribution of the different article categories to the toponym relationships.

We also received specific feedback on Task 4. Our initial idea was to depict small multiple maps with different centuries next to one another as an alternative approach to the network interface with drop-down menus and time sliders. However, one participant in the *think aloud study* raised concerns regarding the visualization of large networks next to one another on a web page due to the display space limitation. Only a few toponyms might be displayable in such small multiple networks which reduces the *utility* of such an interface. We therefore decided against this alternative network interface approach and focused on Tasks 1, 2, and 3 instead. We further decided to exclude Task 6 for future prototype implementations. This decision was based on the (technical) complexity of the required functionalities. Two participants had already expressed concerns related to understanding the interface components in the *think aloud sessions*, which were related to the complex presentation of the content and the information in the spatialized display.

Summary

We presented paper mockups and a list of tasks to target users. We asked participants to comment on their thoughts while solving the tasks and received feedback on potential issues with the web interfaces. We decided to fix interface issues which we ranked low in *difficulty*, but high in *importance*. Most issues we identified are related to user interface elements (e.g., zoom functionality).

As a result of the *think aloud study*, we revised the interface concept. This is described in the following subsection.

5.3.2 Prototype implementation

In this subsection, we illustrate the prototype implementation of the *spatialized network interface* and the *thematic landscape*. The interface design choices were based on the results of the empirical evaluations presented in *Subsection 5.3.1*. Both prototypes incorporate *distant* and *close reading* functionalities (Moretti, 2005, Jockers, 2013) and are designed based on Shneiderman's (1996) *visual information-seeking mantra*: "*overview* first, *zoom* and *filter*, then *details-on-demand*" which was introduced in *Subsection 4.3.2*. The prototypes aim at providing interested information seekers with exploratory and interactive access to spatio-temporal and thematic information in the HDS.

We first introduce the interactive *spatialized network interface* and then present the *thematic landscape*.

Spatialized network interface

In Figure 47, the *spatialized network interface* of the *19th century* at the spatial scale *Switzerland* is depicted as an example. Numbered areas highlighted in grey represent interaction elements of the interface and are described here. Changes in the interface concept and design of the prototype interface, as a response to the evaluation results in *Subsection 5.3.1*, are highlighted in this subsection.

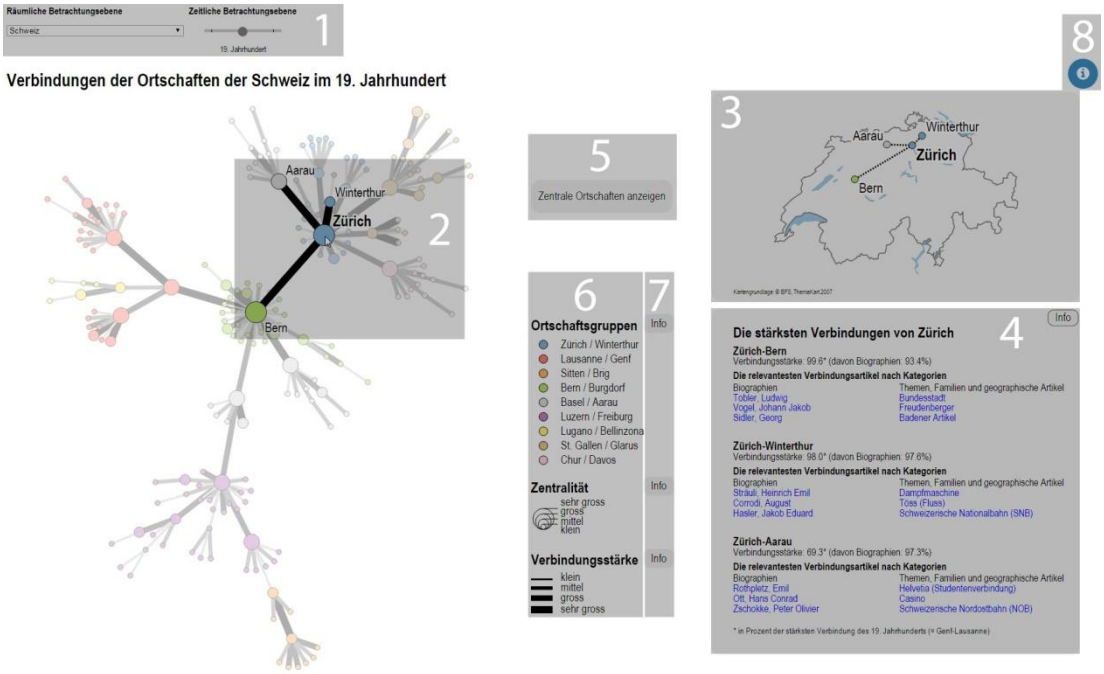


Figure 47: Interface elements for the *spatialized network display*.

The selection of a spatial scale is implemented with a drop-down menu, while the selection of a temporal scale is implemented with a time slider (number 1 in Figure 47). Both the drop-down menu and the time slider are depicted above the network visualization, which was suggested by a participant of the *think aloud study* (see *Subsection 5.3.1*). A user can interact with the network visualization by doing a mouseover or clicking. In Figure 47, a user has selected the toponym *Zürich* by mouseover. As a result of this, the three strongest toponym relationships of *Zürich* are displayed and the toponyms are labeled (number 2 in Figure 47). In addition, the relationships are depicted and the toponyms are labeled in the map (number 3 in Figure 47). This dynamic labeling functionality was implemented based on feedback obtained from participants who took part in the *think aloud study* (see *Subsection 5.3.1*). The selected toponym (i.e., *Zürich*) is labeled in bold, while the other toponyms are labeled in regular font.

When a user clicks on a toponym in the *spatialized network interface*, it results in an altered view of the info window (number 4 in Figure 47). This part of the interface is depicted in an enlarged view in Figure 48 for the toponym *Zürich*. The three strongest toponym relationships and their strength (= *Verbindungsstärke*) are displayed. The relationship strength is normalized to the strongest existing relationship (= 100) in the respective century. Overall, the strongest relationship in the 19th century is *Genf-Lausanne* (100%).

The strength of *Zürich-Bern* is 99.6, which implies that the strength of the toponym relationship *Zürich-Bern* is nearly as strong as *Genf-Lausanne*. The contribution of the article category *biographies* to each toponym relationship is displayed in brackets (= *davon Biographien*). Further below, the titles of the most relevant articles (= *Die relevantesten Verbindungsartikel nach Kategorien*) according to the weighted spatio-temporal relationship algorithm (see *Subsection 4.2.1*) are displayed for each relationship. The articles are displayed in two columns: in the left column, *biographies* are listed (= *Biographien*); in the right column, *thematic contributions* and articles about *families* and *geographical entities* (= *Themen, Familien und geographische Artikel*) are displayed. We decided to split the info window into these two columns, as *biographies* dominate the shown toponym relationships (see *Subsection 5.2.1*), and for many relationships, only *biographies* would have been displayed if only the strongest three articles were shown. Hyperlinked article titles represent an additional functionality. When a user clicks an article title in the information window, a new tab is opened in the browser to display the article directly on the HDS website.

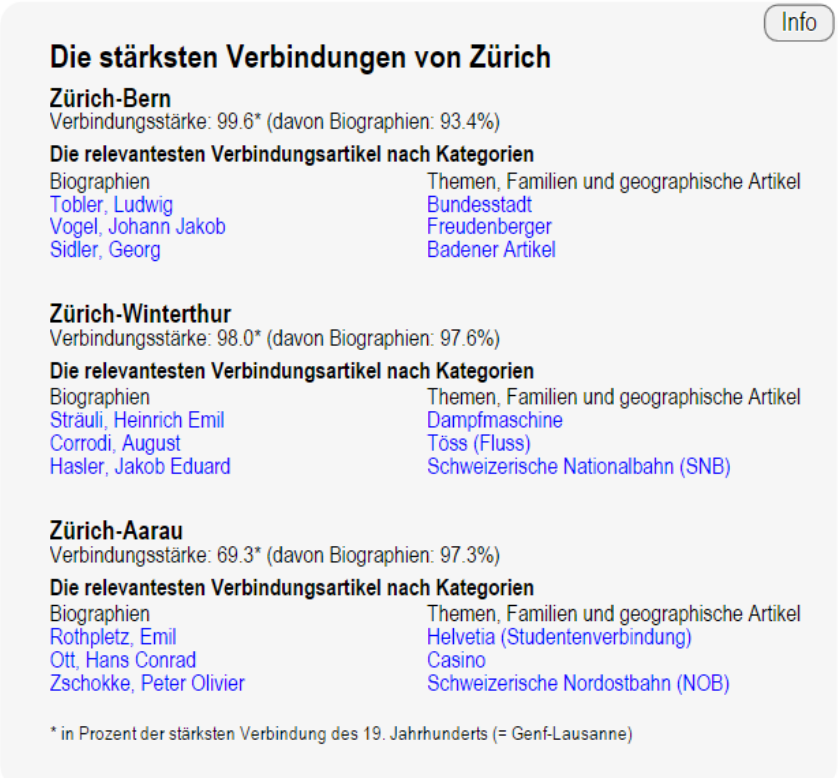


Figure 48: Info window of the toponym *Zürich*.

If a user clicks on an edge in the network visualization, the same content as in Figure 48 is displayed in the information window, but only for the selected toponym relationship. The interaction modalities to display information in the info window, and the actual design of the info window, are adapted from Figure 45. Instead of displaying five toponym relationships, we decided to only include the strongest three toponym relationships which was suggested in the *focus group meeting*. Then, we decided to include articles instead of article categories in the info window, which was suggested by a *think aloud study* participant and because participants in the *focus group* found it important

to be able to access the source information. This allows a user to directly access these HDS articles in which two toponyms often co-occur.

The *spatialized network interface* further incorporates a functionality to highlight the most central toponyms (number 5 in Figure 47). By placing the mouse cursor in the *Zentrale Ortschaften anzeigen* field in Figure 47, all toponyms classified as *highly* or *very highly* central are highlighted and labeled in the network visualization and in the map. This functionality is included in order to provide users an idea of where most central toponyms are located in the network and in the map, which might be particularly helpful for users who are unfamiliar with the overall structure of the network or the toponyms of Switzerland. In addition, a legend of the network visualization is displayed in Figure 47 (number 6) and info boxes are shown (number 7). The info boxes were created as a response to feedback in the *think aloud study*. They contain further information regarding the legend. Another info box is displayed in the top right corner of the info window in Figure 48. It briefly explains the content of the info window. A help page further explains technical aspects of the web interface. It is opened once a user clicks on the blue button in the top right corner of the interface (number 8 in Figure 47). In addition to technical instructions, the help page contains a link to the *project description* web page. This page contains information about the project, the aims of the project, the data source, the computation and visualization, as well as publications and literature links. This page was created in response to the *transparency* requirement suggested during the *focus group* (see *Subsection 5.3.1*).

At this stage of the design process, the direct selection of toponyms in the map is not included. This might be a useful feature to be included in a full release of the interface.

So far, we illustrated how spatio-temporal interconnections, retrieved from the HDS corpus, can be explored through an interactive *spatialized network interface*. In the next step, we illustrate how we incorporated the thematic data in an interactive *thematic landscape*.

Thematic landscape

The *thematic landscape* interface contains two semantic zoom levels: an *overview* and a *detail view*. The visualization and the incorporation of the two zoom levels in the interactive *thematic landscape* interface are described here. In Figure 49, the *thematic landscape* at the *overview* zoom level is visualized. Specific interface elements are numbered and described in the following paragraphs. Changes to the interface design in response to user evaluation (see *Subsection 5.3.1*) are highlighted in this subsection.

The *thematic landscape* allows users to zoom (number 1 in Figure 49), following Shneiderman's (1996) *visual information-seeking mantra*: "*overview* first, *zoom* and *filter*, then *details-on-demand*". Standard zoom tools are used, operating in ways identical to commonly used geobrowsing tools such as *Google Maps*. We decided to incorporate this functionality as participants in the *think aloud studies* reported disorientation with zooming tools unfamiliar to them. Furthermore, an article title search tool with an

auto-completion function is incorporated (number 2 in Figure 49). Therefore, users can search for a specific article in the *thematic landscape*. In addition, a dynamic legend is displayed by clicking on the blue button at the bottom of the map (number 3 in Figure 49). This window explains the colors used in the *thematic landscape*. This is differently implemented than the design in the paper mockups (see *Subsection 5.3.1*). We initially planned to visualize a static legend next to the interface. However, we decided to let users choose whether they wished to look at the legend as it contains many more elements (i.e., 28 *themes*) than first expected when we ran empirical evaluations. We realized that a static legend including that many elements may distract users from looking at other contents of the interface. Similarly to the *spatialized network interface* previously described, we incorporated a help page in the top right corner of the interface (number 4 in Figure 49). By clicking on the blue button, technical instructions are displayed, and a link to the *project description* web page (i.e., the same web page as for the *spatialized network interface*) is provided.

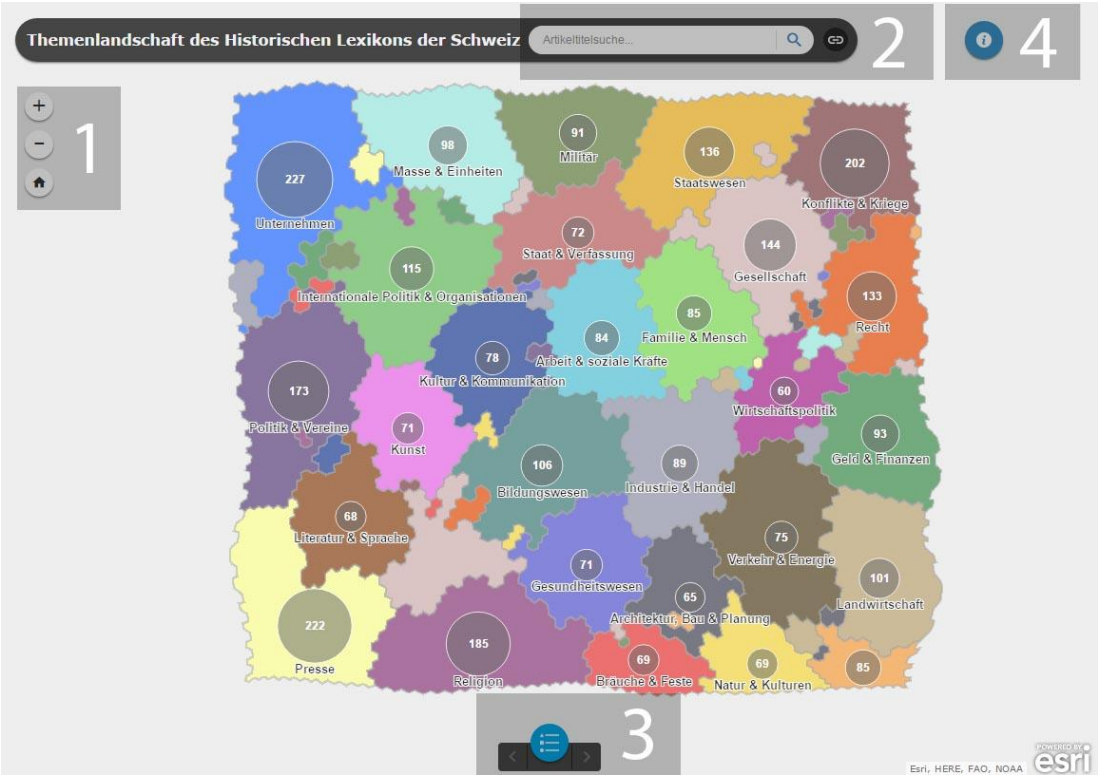


Figure 49: Interaction elements of the *thematic landscape* in the *overview*.

The visualization of the *themes* in the landscape is equal to that of Figure 42. However, the labels of the *themes* are in German (i.e., the chosen interface language). In addition, the size of the *themes* (i.e., number of articles in each theme) is visualized by the size of the circle for each *theme* and a number (e.g., 227 for the theme *Unternehmen* in the top left corner of the map). The larger the circle area, the more articles are contained in that *themes*' region.

We chose only one hierarchical level of *themes* (i.e., no *sub-themes*), which is different from the paper mockup shown in Figure 46. This is due to the large number of *themes* which

were automatically created. We initially expected only a few *themes* at the highest level (i.e., eight *themes* as shown in Figure 46). However, as the automatic clustering resulted in 28 *themes*, we decided not to further divide the *themes* into *sub-themes*.

We depicted the *detail view* of the *thematic landscape* in Figure 50. This view is shown when users zoom in. In this example, we zoomed into the *Companies* (= *Unternehmen*) region in the top left corner of the *thematic landscape*, and searched for *Nestlé* articles in the article query box (number 1 in Figure 50). When a user clicks on the *Nestlé* article in the search tool or directly selects it in the *thematic landscape*, a pop-up window appears containing the article title and titles of the 10 thematically most similar articles. All article titles are implemented as hyperlinks, allowing a user to access the original articles on the HDS website. This functionality was already planned in the *cognitive walkthrough* and *think aloud sessions* and is therefore also illustrated in the paper mockups in Figure 46. However, in contrast to Figure 46, we not only included the three thematically most similar articles. Instead, we included the ten thematically most similar articles in order to provide users with more detailed information.

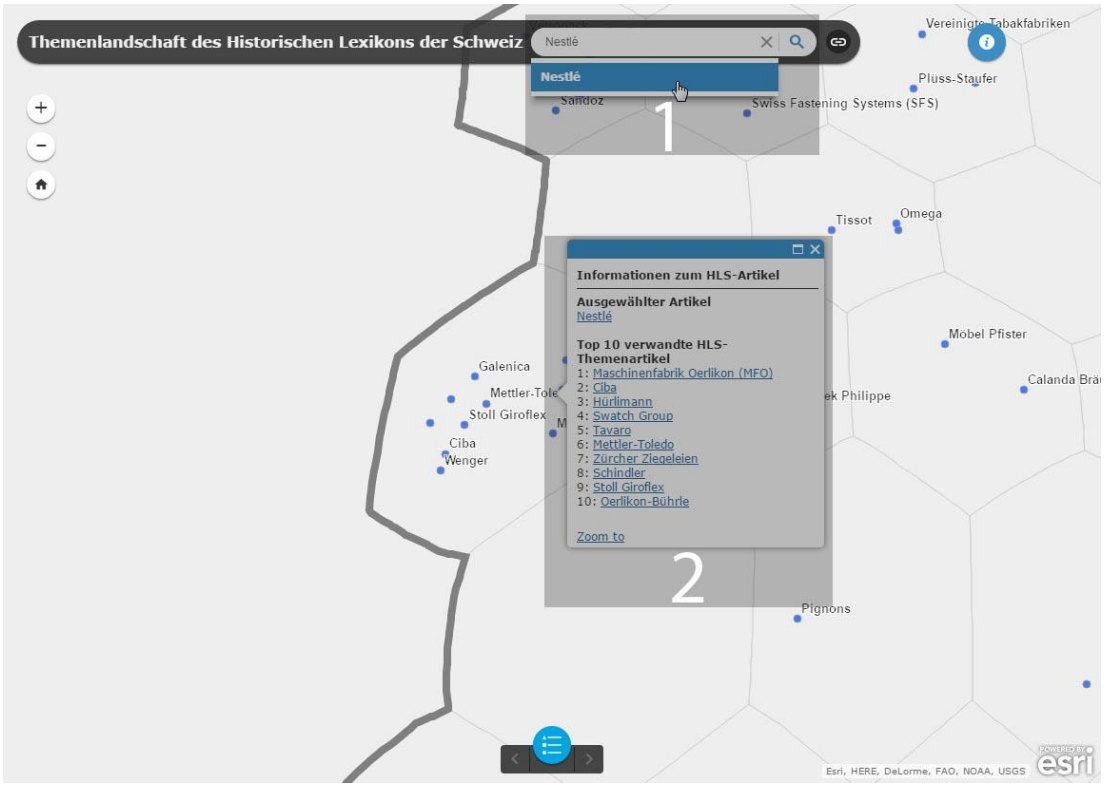


Figure 50: Interaction elements of the *thematic landscape* in the *detail view*.

In contrast to the mockup in Figure 46, the inset map to illustrate the context of the zoomed in region in the detail view is absent in the *thematic landscape*. This is due to current technical limitations of the applied software *ArcGIS Online*. Inset map functionality would be available directly, but only for geographically referenced maps. We might consider an inset map for the *thematic landscape* in a future release of the interface based on the potential future developments of *ArcGIS Online* that might suit our goals.

In this subsection, we have illustrated the two prototype implementations we developed for this project. In the next step, we presented these prototypes to our target users and evaluated them empirically. The results of this empirical evaluation are illustrated in the following subsection.

5.3.3 Empirical evaluation of the prototype implementation

We evaluated the prototype implementation in a combined *utility* and *usability* study, again applying the *think aloud method*. The *usability* evaluation of the design was not our primary aim. Instead, our aim was to gain insight into the types of data exploration users would be able to generate, which includes the users' *knowledge discovery* process (as detailed in *Subsection 4.3.3*).

In this section, we first detail the results of the *spatialized network interface* evaluation and then turn to the *thematic landscape*. For both evaluations, we illustrate the tasks presented to the users, detail how much time users spent with specific parts of the interfaces, and report the results of the *utility* and *usability* evaluation.

Spatialized network interface

To test the prototype implementation of the *spatialized network display*, we asked participants to solve an open task with an online version of the tool (as described in *Subsection 4.3.3*). The reason for choosing an open task is that we were particularly interested in determining which insights are gained by the participants who freely interact with the interface, as opposed to guiding the information-seeking process by providing a very specific task. The task was translated from German and is as follows.

> **Spatialized network interface task**
>
> Open the *spatialized network display*. Use this tool following your own interests. Comment on your display interactions and insights as well as how you gained these insights. As soon as you think that you found all insights which are interesting to you, the task is solved. You can spend a maximum of 40 minutes on this task.

We also prepared sample questions and asked participants to only consider these if they should need it to carry on with their discovery process. These sample questions were as follows.

- What are interesting toponym relationships in a specific century? Which are the most relevant articles for these toponym relationships (i.e., *Verbindungsartikel* in Figure 48)? Comment on your insights, and tell us why you expected these or why these are surprising to you.

- Does the arrangement of connected toponyms in the network correspond with their spatial arrangements in the map? How do the arrangements in the network and map relate to the *toponym communities* (i.e., *Ortschaftsgruppen* in Figure 47)? How do you explain the insights you gained?

- Which toponyms are central to the network of a specific century and to which toponyms are they connected? How does this pattern evolve over time? How do you explain the insights you gained?

We recorded participants' interactions with the tool automatically and audio-taped the *think aloud session* as described in *Subsection 4.3.3*. The interactions and audio records were then used to determine how long participants interacted with the different parts of the interface, and to analyze which insights they gained.

Results – Duration

Participants completed the study after approximately 38:20 minutes, on average. Two out of five participants worked for the maximum allowed 40 minutes on the task. Figure 51 depicts a timeline displaying when and for how long participants explored the six available *spatialized network displays* (represented by colors). The horizontal axis represents the study time with a temporal resolution of 15 seconds. On the vertical axis, the five participants are listed.

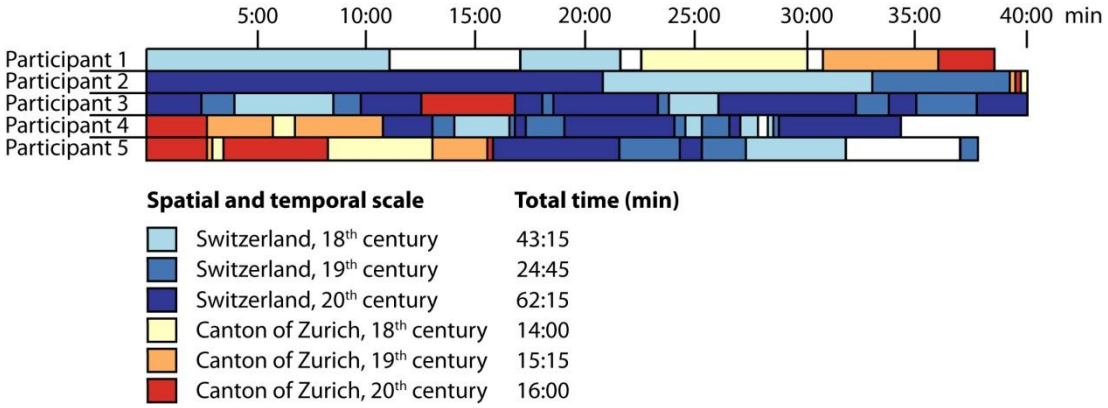


Figure 51: Time spent on *spatialized network displays* of different spatial and temporal scales.

As Figure 51 illustrates, participants spent most of their time viewing *spatialized network displays* of Switzerland (i.e., 130:15 min). For approximately ¼ of their time (i.e., 45:15 min), participants viewed the networks about the *Canton of Zurich*. The network visualizations about the 20th century were used the longest for both spatial scales. The *Canton of Zurich* in the 18th century was the least used network display (i.e., 14 min). In contrast, participants worked for over 40 minutes on the network about *Switzerland* in the 18th century. In Figure 51, gaps in white indicate that participants were not actively interacting with the interface during this time, but looked up other project relevant information, read through the task again, or were expressing ideas for further functionalities of the web interface.

The pattern in Figure 51 clearly shows that all five participants used the interface differently. Whereas Participant 1 only switched three times between four different networks, Participant 4 switched 19 times between six networks. Another noteworthy point is that Participant 4 only interacted for 34:15 minutes with the interface, thus 4:25 minutes less than Participant 1 and the shortest overall. Participant 2 worked for the longest consecutive time period on a single network display. Specifically, this participant interacted with the network *Switzerland, 20th century* for 20:45 minutes in total.

We expected that participants spend more time on the networks of *Switzerland* than on the networks of the *Canton of Zurich* because the networks of the *Canton of Zurich* only contain few nodes and thus the potential to gain insights is limited compared to the large networks of *Switzerland*. Participants confirmed this expectation. Furthermore, participants mentioned during the *think aloud sessions* that they wished to see more *thematic contributions* and articles about *families* and *geographical entities* for the toponym relationships in the *Canton of Zurich* networks instead of almost only *biographies*, which is due to the fact that the *biographies* article category contributes to more than 96% of the networks for the *Canton of Zurich*, as shown in Figure 41.

Results – Utility

Next, we detail the results obtained by analyzing the *participants' interaction* and *audio records*. As we focused on gained insights, we first extracted insights that participants either stated explicitly (i.e., "by interacting with this spatialized display, I found out that…") or implicitly (i.e., "I see that…") in the *think aloud sessions*. Then we summarized these insights in a list for each participant. In the next step, all five lists were consolidated into one combined list. Duplicates (i.e., two people generating the same insight) were removed. In the resulting list, we rated each insight for *complexity*, *depth*, *unexpectedness*, and *relevance*, based on North's (2006) recommendations (see *Subsection 4.3.3*). The more *complex*, *deep*, *unexpected*, and *relevant* the insight, the higher the rating for *complexity*, *depth*, *unexpectedness*, and *relevance*, respectively, on a five point *Likert scale*.

Complexity, *depth*, and *unexpectedness* were rated by the author of this thesis by analyzing the *participants' interaction* and the *audio records*. To rate *complexity*, we analyzed how much data the participants incorporated to gain an insight, following North (2006). For example, if a participant considered toponym relationships in many different centuries to generate an insight, we rated it as a complex insight. To rate the *depth* of an insight, we analyzed if the generation of the insight included many steps, as suggested by North (2006). For example, if a participant considered and combined many different aspects to formulate an insight, instead of stating a simple obvious fact, we rated it as a deep insight. In order to rate the *unexpectedness*, we used the audio records and analyzed if participants either explicitly or implicitly (e.g., by their tone of voice) stated if they expected the insight they gained or not. In order to analyze the *relevance*, an expert historian was involved. The historian rated an insight as relevant if it is deeply embedded in the data domain and connects data to existing domain knowledge.

Most interesting insights score high in all the four categories previously listed, according to North (2006). We list the highest overall ranking insights in Table 18. We decided to

incorporate six out of 32 total insights in Table 18, as these six insights are representative of all insights gained by the participants. For each category (i.e., *complex*, *deep*, *unexpected*, and *relevant*) ratings are shown in Table 18. For example, a value of 1 in the column labeled *complex* would imply that an insight is very simple, whereas a value of 5 would imply that an insight is very complex. In the rightmost column of Table 18, the total score of all rating categories is shown. A complete list of insights is presented in Appendix C in German (i.e., the chosen interface and *think aloud study* language). In the following text, we analyze the six insights in greater detail. We focus on how insights were gained and which information participants combined in order to gain the insights.

Table 18: Insights and *complexity*, *depth*, *unexpectedness*, and *relevance* ratings.

Insight	Complex	Deep	Unexpected	Relevant	Total
1) Although *Fischingen* and *Basel* are geographically distant from one another, they are strongly connected in the 20th century due to the topic *religion*.	3	4	5	5	17
2) The position of *Zürich* in the 18th century network of Switzerland is not central and *Zürich* and *Bern* are not directly connected. In the 19th century and in the 20th century *Zürich* and *Bern* are directly connected and Zürich becomes a very central node in the network.	5	4	3	5	17
3) The *Lower Valais* (e.g., Monthey) and the *Upper Valais* (e.g., Ernen) are connected in the 18th century due to territory in the *Lower Valais* which was owned by a *bailiff* (= *Landvogt*) of the *Upper Valais*.	2	4	5	5	16
4) The network structure of the 18th century at the spatial scale *Switzerland* is linear, whereas the 19th and particularly the 20th century network of Switzerland is more centralized and dominated by central places.	5	4	2	5	16
5) *Lausanne* and *Genf* are directly connected to *Bern* in the 19th century, whereas in the 20th century they are directly connected to *Zürich*.	4	3	4	5	16
6) Confessional relationships between toponyms dominate the networks and sometimes influence the toponym relationships more than geographic distances. For example, *Stans* and *Appenzell* in the 20th century, and *Einsiedeln* and *St. Gallen* in the 19th century are visualized close together in the network, despite of a large geographic distance between them.	3	5	3	5	16

All of the participants' insights listed in Table 18 were rated highly relevant (i.e., 5 in the *relevant* column) by the expert historian, and all six insights have a total score of either 16 or 17 out of 20. The first insight that is reported in Table 18 was gained as participants

interacted with both the network visualization and the map. They detected that, although, the two toponyms *Basel* and *Fischingen* are geographically distant from one another, they are located in close proximity in the network visualization of Switzerland in the 20th century. Participants read articles which were listed in the info window (e.g., number 4 in Figure 47 points to an info window) of the toponym relationship *Basel-Fischingen* and stated that this relationship is particularly based on the topic *religion* (e.g., religious people who lived in one place and studied in the other). Participants mentioned, during the *think aloud sessions*, that they did not expect this toponym relationship to be strong; therefore, the rating for the *unexpectedness* category in Table 18 is 5. Some participants even mentioned that they might follow up this insight after the *think aloud study* in order to find out more about this unexpected relationship.

Insight 6 is thematically related to Insight 1, as it describes that confession has a large influence on the toponym relationships in the network. In contrast to Insight 1, Insight 6 is more general and does not focus on a single toponym relationship. Participants were not very surprised about the strong influence of confession on the toponym relationships in general because confession/religion had a strong impact on Swiss history and is documented extensively in the HDS. Therefore, the rating for *unexpectedness* is 3 for Insight 6 in Table 18.

We rated Insight 2 in Table 18 as highly complex (i.e., *complexity* rating of 5) as participants involved a greater amount of data (i.e., networks of different centuries) to generate this insight. Participants discovered that the position of *Zürich* is not central, and that *Zürich* and *Bern* are not directly connected to one another in the 18th century network, whereas, in the 19th and 20th centuries, *Zürich* and *Bern* are directly connected and *Zürich* becomes increasingly central in the network over time. Therefore, participants analyzed the hierarchical structures presented in the networks (i.e., which nodes are most central) and compared structures over time.

Similarly to Insight 2, participants analyzed the hierarchical structure of the networks in different centuries to generate Insight 4. They stated that the network structure of the 18th century is linear, whereas the 19th and particularly the 20th century networks of Switzerland are more centralized. Therefore, central places such as *Zürich* and *Bern* are located in the middle of the network, and many nodes are directly connected to these central nodes. Similarly to these insights, participants analyzed the toponym relationship structure of different centuries in order to generate Insight 5 in Table 18. They revealed that *Lausanne* and *Genf* are directly connected to *Bern* in the 19th century, whereas they are both directly connected to *Zürich* in the 20th century.

Participants also gained further insights related to territorial ownership, which is exemplified by Insight 3. Participants interacted with the network and map of Switzerland in the 18th century and discovered that the *Lower* and the *Upper Valais* (i.e., regions in the south of Switzerland) are connected. Then, participants read articles about the 18th century in which toponyms of the two regions (e.g., *Ernen* and *Monthey*) co-occur and uncovered that this relationship is based on the fact that territory in the *Lower Valais* was owned by a *bailiff* (= *Landvogt*) of the *Upper Valais* at that time.

These six highest-ranking insights were all gained at the country level, which is not surprising because participants worked substantially more at this scale than at the networks of the *Canton of Zurich*, as explained in the beginning of this subsection. The insights for the *Canton of Zurich* are similar to those listed in Table 18. For example, participants revealed that *Zürich* is directly connected to all toponyms which were owned by the city of *Zürich* in the 18th century. Participants were not surprised about this finding, which resulted in a low *unexpectedness* rating. Furthermore, the amount of data which the participants incorporated to gain this insight (i.e., only the 10-nodes network of the *Canton of Zurich* in the 18th century was considered) is very low and thus the *complexity* rating is low. As a result of these low ratings, this insight is not part of the six highest-ranking insights listed in Table 18.

To summarize, we illustrated that participants used both the *distant reading* (i.e., *spatialized network display*) and the *close reading* (i.e., reading HDS articles) approach to gain insights regarding spatio-temporal patterns in the HDS. Many of the insights participants gained in this study confirm our expectations. For example, we illustrated that participants analyzed and compared the hierarchical structure of the *spatialized network displays* (i.e., Insights 2, 4, and 5). This is what we expected, as network visualizations and the layout algorithm we implemented (see *Subsection 4.2.1*) highlight hierarchical structures and relationships. In addition, we were not surprised that *religion* has a strong influence on the networks (i.e., Insights 1 and 6) due to the importance of *religion* to the history of Switzerland. This is supported by Figure 42, as *religion* even forms its own cluster of articles in the *thematic landscape*, and 185 out of 3,067 *thematic contribution* articles are part of this cluster.

In contrast, we were surprised that participants generated highly complex and deep insights by combining much information (e.g., comparing network structures of several centuries to one another) in order to gain new insights (e.g., Insights 2 and 4) during the *think aloud sessions*. We expected that such complex and deep insights are only possible if participants would have more than 40 minutes time to interact with the interface.

Results – Usability

In order to test the *usability* of the interface, participants were asked to fill in a *System Usability Scale (SUS)* questionnaire. The SUS scores range from 0 to 100 and the higher the score, the better the *usability*. Bangor et al. (2008) collected data from 2,324 surveys which analyzed SUS scores and reported a mean value of 70.1. In their work, they suggest a minimum value of 70 for an acceptable system, and that better products regarding *usability* should have a value in the high 70s to upper 80s (Bangor et al., 2008: 592). For the *spatialized network interface*, the participants provided an average score of 78. The SUS score for the *spatialized network interface* thus fulfills Bangor et al.'s (2008) *better products criterion*.

Furthermore, we asked participants whether they are satisfied with the results they obtained, how confident they were that they reached the goal of the task, and how relevant they thought their insights are regarding the history of Switzerland. On average, participants rated their satisfaction with the obtained results 3.8 on a five point

Likert Scale, ranging from very unsatisfied (= 1) to very satisfied (= 5). The question of how confident they are about reaching the goal was rated 4.0 on average, ranging from not confident at all (= 1) to very confident (= 5) on a five point *Likert Scale*. The relevance of their insights for Swiss history was judged 2.6 on average on a five point *Likert Scale*, ranging from not relevant at all (=1) to very relevant (= 5). These numbers show that participants were satisfied about the results they achieved and think that they reached the goal of the task, but did not think the insights they gained were highly relevant for Swiss history, which contradicts the expert judgments of relevance in Table 18, which are very high. This could be explained by the fact that participants did not define relevance in the same way the historian expert and we did for the rating in Table 18. Many participants commented on this question and began discussing the definition of relevance while filling out the questionnaire. Most of the participants interpreted relevance as to which degree the state of the art in Swiss history research was extended, rather than how deeply the insight is embedded in the data domain and connects data to existing domain knowledge, which is how we define relevance to judge insights based on North (2006). This could potentially explain the comparatively low score of 2.6 compared to the high relevance scores in Table 18.

Due to the low number of participants (n=5), no further statistical analyses were computed for the SUS or the *Likert Scale* scores. The scores are qualitatively compared to the *thematic landscape* interface scores in the next section.

Results – Further comments

Participants were invited to optionally provide further feedback regarding the interface. One participant mentioned that the use of the interface might inspire them to think about toponym relationships which are surprising at first glance. However, the same participant stated that the motivation to work with the tool was only high in the beginning. In the last minutes of the study, the level of motivation decreased because the participant did not gain interesting insights anymore, and the participant mentioned most probably not to use the tool again in the future.

In contrast, another participant was excited about the design and provided functionalities for the tool. Furthermore, the participant was interested in incorporating the tool into own projects related to the interactive visualization of history. The same participant mentioned that the interface might be best for historians with a high interest in *biographies*, as this article category contributes the most to the spatio-temporal relationships (as illustrated in *Subsection 5.2.1*).

Having illustrated the evaluation results of the *spatialized network interface*, we now turn to the evaluation results for the interactive *thematic landscape* interface.

Thematic landscape

In order to test the prototype implementation of the *thematic landscape*, we asked participants to solve a more specific task, compared to the *spatialized network interface*. We decided to define a more specific task because we were particularly interested if

participants use the interactive *thematic landscape* to search for thematic information as we expected. We expected that if we ask participants to search for articles about a combination of two neighboring *themes* in the *thematic landscape* (e.g., *Religion* and *Customs & festivals*, see Figure 42) to search articles in the border region of these two *themes* in the *thematic landscape*. If participants would do so, we could assume that they understood the principle that similar articles and *themes* are placed in close proximity in the *thematic landscape* (i.e., articles are placed in border regions if they are related to both *themes*). Therefore, our assumption that a *thematic landscape* might support the thematic information search process would be supported (see *Subsection 5.2.2*). The task was translated from German and is as follows.

Thematic landscape interface task

Open the *thematic landscape tool*. Find five HDS articles which fit best into the fictitious topic *religious customs and festivals*. During the first 5 minutes of the study, you are only allowed to use the *thematic landscape* to solve the task. In the last 10 minutes of the study, in addition to the *thematic landscape*, you can also use the article and text search functionalities provided on the official website of the HDS to solve the task. As soon as you write down the five HDS articles, the task is solved. You are not required to rank the articles. You can spend a maximum of 15 minutes on this task.

We specifically selected the fictitious topic *religious customs and festivals* because we expected participants to search for articles in the border region of the neighboring *Religion* and *Customs & festivals* themes in the *thematic landscape* (see Figure 42). This topic was chosen based on the expectation that all participants are somewhat familiar with this topic. Furthermore, we were interested to see whether participants would use the *thematic landscape*, or if they would use the HDS website to find relevant articles.

Duration

Participants decided to finish the study after approximately 13:45 minutes on average. Two out of five participants worked for the maximum 15 minutes of time allowed for the task. Figure 52 depicts a timeline that shows when participants used the *thematic landscape* interface, and when they accessed and used the HDS website (i.e., the e-HDS) to search for relevant articles. The horizontal axis represents the study time with a temporal resolution of 15 seconds. On the vertical axis, the five participants are listed.

As Figure 52 illustrates, only two out of five participants used the e-HDS search functionalities to solve the task. Participants interacted with the *thematic landscape* interface during a time period of 62:30 minutes, which equals 93% of the total time spent by all of the participants for the study. The remaining 4:45 minutes participants navigated on the HDS website, using the full text search function or the advanced search capabilities of the e-HDS.

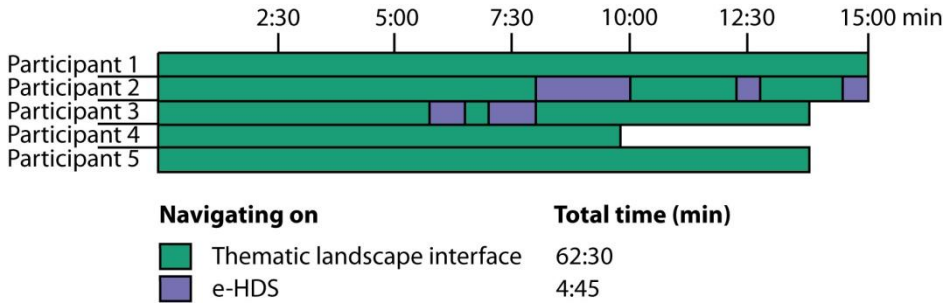


Figure 52: Time spent interacting with the *thematic landscape* interface and the e-HDS.

Results – Utility

In Figure 53, the locations of the articles that participants identified to fit best into the fictitious topic *religious customs and festivals* in the *thematic landscape* are depicted. The *detail view* map (i.e., the zoomed in view which covers the largest part of the figure) in Figure 53 represents an enlarged view for the area framed by a black rectangle in the *overview map* shown in the top right corner of Figure 53. The colors represent the thematic regions in the *thematic landscape* as introduced in Figure 42 (see *Subsection 5.2.2*). In the *overview map*, the themes *Religion* (= R) and *Customs & festivals* (= C&F) are labeled.

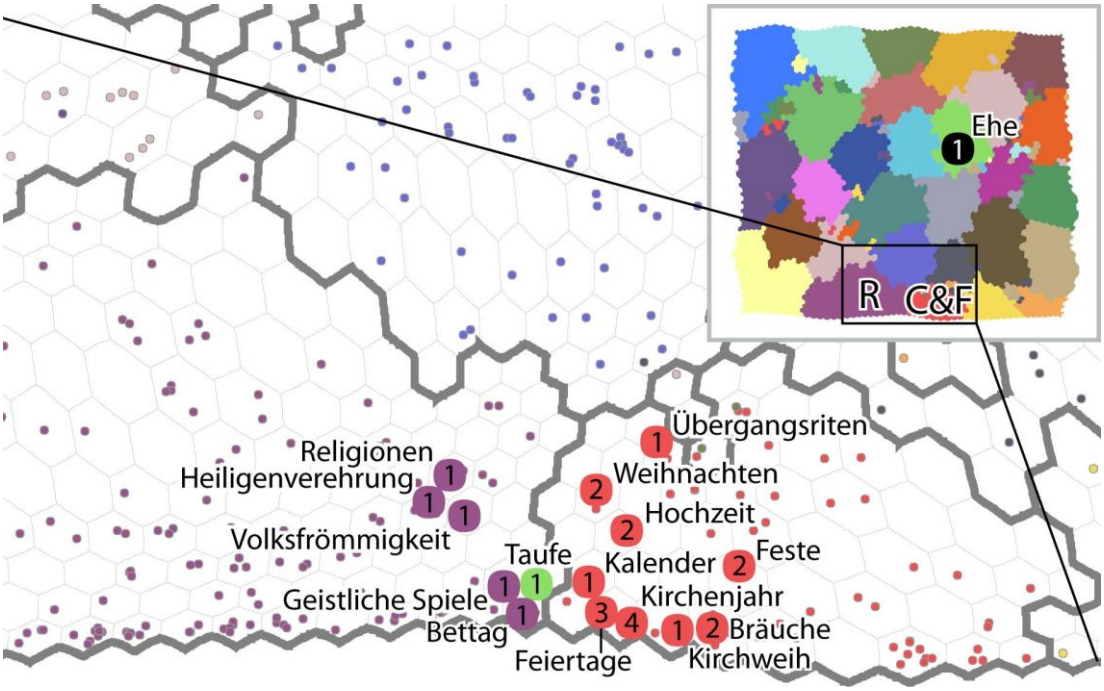


Figure 53: HDS articles that fit best into the topic *religious customs and festivals*.

In the *detail view*, all articles which were identified by at least one participant are visualized as large points. Within these large points, a number indicates how many participants chose the respective articles, all of which are labeled. For example, the article *Weihnachten* (= Christmas), which is located in the center of the *detail view*, was mentioned by two participants, whereas the article *Taufe* (= baptism), at the bottom of

the *detail view*, was only mentioned by one participant. All other articles in the *detail view* of the *thematic landscape* in Figure 53 are visualized as small and unlabeled points. The colors represent the *themes* in the *thematic landscape*, as shown in Figure 42.

One participant mentioned the article *Ehe* (= marriage) as fitting into the fictitious topic *religious customs and festivals*. However, this article is not located in the region of the violet *Religion* or the red *Customs & festivals* theme in the *thematic landscape*, but in the green *Family & humans* theme. Therefore, this article is not shown in the *detail view*, but in the *overview map* in the top right corner of Figure 53.

As shown in Figure 53, 24 out of 25 articles chosen by participants are indeed located in the border region of the two themes *Religion* and *Customs & festivals*, as expected. All participants followed a similar strategy by using the zooming and panning functionalities to zoom to the *detail view* of the border region between the *Religion* and *Customs & festivals* themes and then began to read the displayed article titles. Participants opened the pop-up windows of potentially relevant articles, and for some of them, they accessed the same article in the e-HDS to check whether the article text was related to the topic *religious customs and festivals*. Furthermore, four out of five participants opened one of the 10 thematically most related articles in the e-HDS at least once. These thematically related articles are displayed in the pop-up windows of each article in the *thematic landscape*. The article search tool in the *thematic landscape* interface was used by all participants at least once.

The two participants who accessed the e-HDS in order to search for articles entered search terms such as *Religion*, *Brauchtum* (= customs), and *Fest* (= festival) to search for relevant articles either with the full text or the advanced search tool capabilities of the e-HDS. The articles *Ehe* (= marriage), *Religionen* (= religions), and *Heiligenverehrung* (= adoration of saints) were found and selected by these two participants. The remaining 22 articles were found by interacting with the *thematic landscape* interface only. A major reason why only few participants used the e-HDS might be that only text queries are possible in the e-HDS, and no query functionalities are provided to search for articles about a specific *theme* (i.e., a group of thematically similar articles) such as *religion* in the *thematic landscape*.

Results – Usability

As detailed earlier for the *spatialized network interface*, participants completed a *System Usability Scale (SUS)* questionnaire in order to test the *usability* of the interface. Participants rated the *thematic landscape*, on average, with a value of 81. This value exceeds the minimum value of 70 used to rate an acceptable system, and can be ranked as *better product*, as defined by Bangor et al. (2008: 592). The average score thus indicates an even higher *usability* than for the *spatialized network interface*.

Additionally, we asked participants whether they were satisfied with the results they obtained, how confident they were that they reached the goal of the task, and how relevant they thought their insights are regarding the history of Switzerland. On average, participants rated their satisfaction with the obtained results 4.4 on a five point

Likert Scale, ranging from very unsatisfied (= 1) to very satisfied (= 5). The question of how confident they are about reaching the goal was rated 4.0 on average, ranging from not confident at all (= 1) to very confident (= 5) on a five point *Likert Scale*. The relevance of their insights for Swiss history was judged 2.5 on average on a five point *Likert Scale*, ranging from not relevant at all (=1) to very relevant (= 5). Due to the low number of participants (n=5), no further statistical analyses were computed for the SUS or the *Likert Scale* scores. Therefore, the satisfaction of the participants with the results they achieved is higher compared to the *spatialized network interface*, whereas the confidence level and the relevance judgments are similar. For the relatively low rating regarding relevance, the same explanation as for the *spatialized network interface* can be used and is therefore not repeated here. Due to the low number of participants (n=5), no further statistical numbers or tests were computed for the SUS or the *Likert Scale* scores.

Further comments

Participants were invited to provide further feedback regarding the interface after completing the study. Two participants thought that the interface provides a nice alternative to access the HDS, and that it might serve the HDS to motivate more people to access their dictionary. Another participant added that it is particularly helpful to learn about a topic which is unfamiliar to a potential user. Additionally, two participants mentioned the idea of using the tool at schools. Pupils might be motivated to learn a new history topic with such an interactive and exploratory tool.

One participant mentioned that a topic related to the *Companies* theme would have been more interesting to study, instead of the chosen *Religion* and *Custom & festivals* themes. In contrast, another participant mentioned that the chosen task and themes were very interesting.

To summarize, in this chapter we have illustrated the results we obtained by retrieving spatial, temporal, and thematic information from the HDS (i.e., answering *Research Question 1*), spatializing them in *network visualizations* and a *self-organizing map* (i.e., answering *Research Question 2*), and incorporating them in interactive web interfaces (i.e., answering *Research Question 3*). We involved target users early on in the user interface design process and evaluated two prototype implementations. In the following chapter, we evaluate the *quality* and *sensitivity* of the obtained results.

6 Evaluation

In this chapter, we report on different evaluations to test the *quality* and *sensitivity* of our results to the methods and the parameters applied in this project. In *Section 6.1*, we assess the precision of GIR results in *spatialized networks* on three different levels. In *Section 6.2*, we detail the *sensitivity* of our parameter choices to the *spatialized networks*. In *Section 6.3*, we review the process of selecting a reasonable number of topics for the thematic clustering of HDS articles. In *Section 6.4*, we evaluate the themes assigned to the *thematic contributions* articles of the HDS.

6.1 Precision of GIR in spatialized networks

In this section, we illustrate the evaluation of the proposed *spatialized networks* presented in *Subsection 5.2.1*. We investigate the *sensitivity* of the *spatialized networks* to the GIR approach applied in this thesis. In particular, we analyze the effect of toponyms and temporal references which were incorrectly retrieved from the HDS or incorrectly disambiguated by the GIR algorithm on the toponym relationships in the networks. As we focused on network structure in this evaluation, we chose to systematically assess the *spatialized network* of the 19th century at the country level because this network has a relatively balanced structure compared to the 20th century network, which is strongly centralized, and the 18th network, which represents the least centralized network presented in *Subsection 5.2.1*.

Assessing the spatio-temporal relevance of articles

In IR, documents are considered relevant, if they contain information which corresponds with the information needs of a user (Manning et al., 2009a). Translated to our project, this implies that HDS articles which are displayed to a user in the *spatialized network interface* (see *Subsection 5.3.2*) should contain spatial and temporal information which corresponds to the toponym relationship that a user has selected. We illustrate this with an example: if a user selects the toponym relationship *Bern-Zürich* in the *spatialized network interface* (i.e., by clicking on the edge which connects *Bern* and *Zürich*) to learn about this relationship in the context of the 19th century, HDS articles for the toponym relationship *Bern-Zürich* are displayed in the info window of the interface. If these articles indeed contain information regarding *Bern* and *Zürich* and are

particularly about the 19th century, they are considered relevant. Therefore, articles are relevant if they correspond to our definition of spatio-temporal relationships (presented in *Subsection 4.2.1*).

We analyzed all 197 toponym relationships occurring in the 19th century network (see Figure 39). For each of the toponym relationships, the author of this thesis read all HDS articles which are listed in the respective info windows of the *spatialized network interface*. In Figure 54, the info windows for the toponym relationships *Bern-Zürich* and *Gersau-Schwyz* are displayed. The author of this thesis checked whether the spatial and temporal information contained in each article fulfills the criteria of toponym relationships shown in *Subsection 4.2.1*. In order to evaluate spatial information, we defined that at least 50% of the occurrences of two toponyms in a HDS article need to have been correctly retrieved by the GIR algorithm in order to be considered relevant. For example, in the article *Tobler, Ludwig* (see Figure 54), for at least 50% of the occurrences of the term *Zürich* in the article, the term *Zürich* must refer to the city of *Zürich*. In addition, for at least 50% of the occurrences of the term *Bern*, the term *Bern* must refer to the city of *Bern*. These conditions are met for *Bern* and *Zürich* in the article *Tobler, Ludwig* and thus the article is considered relevant for the toponym relationship *Bern-Zürich* with regard to its spatial information. In contrast, for example, if for more than 50% of the occurrences of the term *Zürich*, the term *Zürich* referred to the *Canton* or the *District* instead of the city of *Zürich*, this article would have been considered irrelevant for the toponym relationship *Bern-Zürich*. The *Tobler, Ludwig* article would also have been considered irrelevant if more than 50% of the occurrences of the term *Bern* did not refer to the city of *Bern*.

In order to evaluate the temporal information, we defined that at least 50% of all temporal references (e.g., *September 27, 1816*) in the HDS articles have to be about the 19th century in order to be considered relevant for a toponym relationship regarding temporal information. If less than 50% of the temporal references in an article are related to the 19th century, the article is considered irrelevant for the toponym relationship.

To summarize, we applied *close reading* (i.e., reading the HDS articles) to determine whether the HDS articles for the 197 toponym relationship in the 19th century are relevant. These relevance judgments were used in order to calculate precision, which is discussed in the following paragraphs.

Calculating the precision of the spatio-temporal GIR results

We decided to calculate precision because we wished to evaluate how many of the HDS articles are relevant out of all HDS articles in the 197 toponym relationships of the 19th century network. An article is considered relevant if both the spatial and temporal information in the article are considered relevant to the toponym relationship. Based on the assessment of relevance at the *article level*, we analyzed the precision on the *toponym relationship* and the *network level* as well. We describe the calculation of precision at all three levels using Figure 54.

In Figure 54, the information windows for *Bern-Zürich* and *Gersau-Schwyz* are displayed. To the right of the article titles, check marks or crosses are displayed. A green check mark implies that the article was evaluated as *relevant* (i.e., both spatial and temporal information is considered relevant for the toponym relationship), while the red cross implies that the article was evaluated as irrelevant.

Precision at the *article level* expresses the fraction of relevant HDS articles as a proportion of articles retrieved in the 197 information windows of toponym relationships in the 19th century. In Figure 54, two information windows are displayed as an example. Seven out of nine articles are shown to be relevant. Therefore, precision at the *article level* is 77.8%. Precision at the *toponym relationship level* describes the average precision of all existing toponym relationships. In Figure 54, the precision for *Bern-Zürich* is 83.3%, as five of six articles are relevant. The precision for *Gersau-Schwyz* is only 66.7%, as two of three articles are relevant. The average precision of these two relationships thus is 75%. The precision at the *network level* expresses the fraction of relevant relationships out of all existing relationships in the network. In Figure 54, *Gersau-Schwyz* is not a relevant toponym relationship because only two articles are relevant, and, according to *Subsection 4.2.1*, toponym relationships must be based on at least three articles in order to be relevant. In contrast, *Bern-Zürich* is based on five relevant articles, thereby fulfilling this criterion. Therefore, precision at the *network level* is 50%, as only one of two relationships is relevant.

Figure 54: Evaluating the toponym relationships *Bern-Zürich* and *Gersau-Schwyz*.

In info windows (e.g., info windows of *Bern-Zürich* and *Gersau-Schwyz* in Figure 54), only the three highest ranked *biographies* (left column in Figure 54) and the three highest ranked articles of the *thematic contributions*, *family* and *geographical entities* article categories (right column in Figure 54) based on the algorithm presented in *Subsection 4.2.1* are shown. For toponym relationships containing less than three relevant articles in the info window (i.e., *Gersau-Schwyz* in Figure 54 and *Freiburg-Rheinau* in Figure 55), we checked

whether other articles are relevant for the toponym relationship according to the toponym relationship definition (presented in *Subsection 4.2.1*), but are not displayed in the info window because they are not among the top three ranked articles for either the *biographies* (i.e., left column in Figures 54 and 55) or the *other article categories* (i.e., right column in Figures 54 and 55). For example, in the toponym relationship *Freiburg-Rheinau* in Figure 55, only two relevant HDS articles are displayed in the info window. However, three other *biographies* relevant to the toponym relationship *Freiburg-Rheinau* exist, but they are not shown in Figure 55 because they are not among the three highest ranked *biographies*. For the toponym relationship *Gersau-Schwyz*, we did not locate any relevant articles other than those displayed in the info window in Figure 54. We considered this point for calculating the precision at the *network level*: the toponym relationship *Freiburg-Rheinau* contains five relevant articles in total (i.e., two relevant articles displayed in the info window plus three relevant articles not displayed) and is thus considered relevant at the *network level*, while the toponym relationship *Gersau-Schwyz* contains only two relevant articles and is thus considered irrelevant at the *network level*.

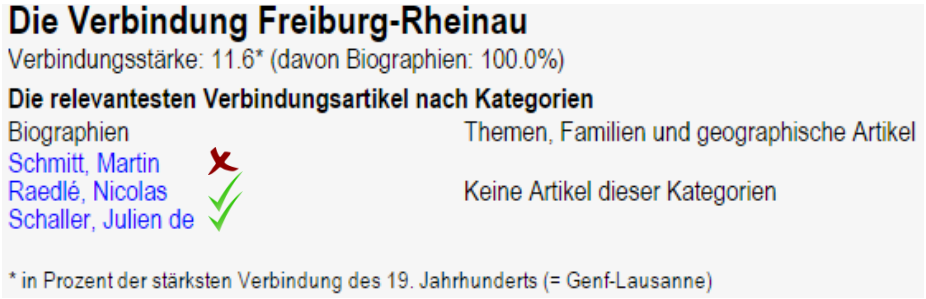


Figure 55: Evaluating the toponym relationship *Freiburg-Rheinau*.

Testing the replicability of the approach

In the next step, we assessed the *interannotator agreement* to test the replicability of the proposed evaluation approach. To do so, the relevance judgments for 50 articles, randomly selected by the author of this thesis, are compared with the relevance judgments of the same articles by two doctoral students at the Department of Geography of the University of Zurich, who are not directly involved in this project. We first explained the 50% rule for spatial and temporal information to judge the relevance of the articles to annotators. Furthermore, we explained additional rules such as the exclusion of the *places of citizenship* from *biographies* (see *Subsection 4.1.1*). Hence, we asked them not to include *places of citizenship* in their relevance judgments. Based on these rules, the two annotators judged each of the 50 randomly chosen articles for whether the article is relevant or not for a given toponym relationship. For example, *Bundesstadt* for the toponym relationship *Bern-Zürich* (see Figure 54) was one of the randomly selected articles. All annotators decided that this article is relevant, because the terms *Bern* and *Zürich* refer mostly to the cities of *Bern* and *Zürich* in the article, and more than 50% of the temporal references in the article refer to the 19th century.

Cohen's kappa (Cohen, 1960) and the *Prevalence and Bias Adjusted Kappa* (PABAK) statistic (Byrt et al., 1993) were used to test the agreement of the relevance judgments of the two

annotators and the relevance judgments of this thesis' author. *Cohen's kappa* measures the chance corrected agreement between two annotators on a nominal or binary scale, and thus expresses the extent to which annotators assigned the same relevance judgments (Tinsley and Weiss, 1975, Banerjee, 1999). The null hypothesis is that no more agreement between two annotators exists than might occur by chance (Gisev et al., 2013: 334). *Cohen's kappa* values for all annotator pairs are depicted in the left half of Table 19. The number of judged articles is 50 for all annotator pairs.

All null hypotheses are rejected ($p < .001$) for all annotator pairs, and thus we determine that the extent of agreement is not caused by chance. The *kappa coefficient* ($= \varkappa$) ranges between 0.56 and 0.81. The averaged *interannotator agreement kappa score* is $\varkappa = 0.69$ (Light, 1971). According to Landis and Koch (1977: 165) this indicates a *substantial* strength of agreement. However, we observed a *prevalence bias*, as annotators judged, on average, 44 of the 50 articles as being relevant and only 6 articles as irrelevant. According to Feinstein and Cicchetti (1990) this *prevalence bias* causes low *kappa scores* even though the *percent agreement* (= number of concordant judgments/number of total judgments) is very high, and thus the *kappa scores* should be corrected. Therefore, we calculated the PABAK statistic, which adjusts *prevalence biased kappa scores* (Byrt et al., 1993). The PABAK statistics for the three annotator pairs are between 0.80 and 0.92 and the average PABAK is 0.87, which corresponds to an *excellent* strength of *interannotator agreement* according to Landis and Koch (1977: 165). The *percent agreement* (PA) and the PABAK for all rater pairs are listed in the right half of the matrix in Table 19.

Table 19: *Cohen's* $\varkappa$ and PABAK.

	Rater 1	**Rater 2**	**Rater 3**
Rater 1	---	PA = 94%	PA = 96%
		PABAK = 0.88	PABAK = 0.92
Rater 2	$\varkappa = 0.69$	---	PA = 90%
	$p < .001$		PABAK = 0.80
Rater 3	$\varkappa = 0.81$	$\varkappa = 0.56$	---
	$p < .001$	$p < .001$	

To summarize, we show that *interannotator agreement* is high and thus the replicability of our manual *close reading* evaluation approach is acceptable. Having illustrated the replicability of our approach, we now turn to the results of the evaluation.

Evaluation results

The 197 toponym relationships in the 19th century network of Switzerland contain 832 articles, which were judged for their relevance for spatial and temporal information by the author of this thesis, as previously mentioned.

Table 20: Precision at *article*, *toponym relationship* and *network level*.

Evaluation level	Precision
Article	83%
Toponym relationship	84%
Network	97%

The precision for the *article*, the *toponym relationship*, and the *network level* are listed in Table 20. Since 687 out of 832 articles were judged relevant, a precision of 83% for *articles* is shown in Table 20. For the *biographies* category, the precision is 92%, whereas the precision for *other article categories* (i.e., *thematic contributions*, *families*, and *geographical entities*) is 61%. This difference is due to the high amount of spatial and temporal information in *biographies* compared to the *other article categories*, as illustrated in *Section 5.1*. The inclusion of up to three articles from *other article categories*, shown in the info windows of the *spatialized network interface* (see Figure 54 and 55, right column), forces the inclusion of low ranked *other article categories* articles according to the toponym relationships algorithm (see *Subsection 4.2.1*). For example, if there are 20 relevant *biographies* and 3 relevant *other article categories* articles for a toponym relationship, only the 3 highest-ranking *biographies* are listed in the info window, whereas for the *other article categories* all articles are considered regardless of their rank because there are only 3 potential articles to choose from. As a result of including such low ranking articles, the precision of articles from *other article categories* is low.

The average precision in *toponym relationships* is 84%, and thus similar to precision for *articles*. Since 6 out of 197 toponym relationships (i.e., 3%) in the 19th century network of Switzerland are irrelevant, precision at the *network level* is 97%. Precision at the *network level* is thus very high and only 6 relationships are incorrectly depicted in the network visualization. To further evaluate incorrectly depicted relationships, we highlighted all irrelevant relationships in the network in red in Figure 56. The nodes of irrelevant toponym relationships are labeled. In addition, we labeled *Zürich*, which is the most central toponym in the 19th century network of Switzerland. All relevant toponym relationships and all toponym nodes are depicted in grey. The size of the nodes represents the strength, and thus the centrality of the toponyms in the network; the larger the node, the higher the centrality of a toponym. As Figure 56 demonstrates, all irrelevant toponym relationships connect peripheral and non-central nodes with other nodes in the network. Toponym relationships between central nodes (e.g., *Bern-Zürich*), the most central connections in the network structure, are all relevant at the *network level* and are therefore not highlighted in red in Figure 56. This implies that the general network structure (i.e., connections between central nodes) for the 19th century is stable regarding the influence of irrelevant spatial and temporal information retrieved from HDS articles.

Figure 56: Irrelevant toponym relationships highlighted in the 19th century network of Switzerland.

Precision at the *article* and *toponym relationship level* presented in Table 20 are similar to those reported by Derungs (2014: 86), at 82%. Derungs (2014) developed the spatial information retrieval algorithm employed in a slightly adapted version in this thesis. Derungs (2014: 86) analyzed a historic corpus of Swiss Alpine literature and evaluated the top 5-6 articles gathered for 10 spatial queries, which is comparable to our approach (we analyzed 3-6 articles per toponym relationship). More recently, Palacio et al. (2015) applied the spatial information retrieval algorithm developed by Derungs (2014) to *Hikr*[75], a user-generated corpus describing mountaineering expeditions. Evaluating the top 10 articles selected from around 4,400 spatial queries resulted in a precision of 70-90%, depending on different spatial search radii (Palacio et al., 2015). Readers interested in the search radii concept and in details related to the evaluation procedure are referred to Palacio et al. (2015). Although our evaluation approach differs in detail from the aforementioned approaches of Derungs (2014) and Palacio et al. (2015) (i.e., we considered spatio-temporal instead of spatial-only queries), the precision we obtain indicates that our combined spatio-temporal approach works equally well. We thus conclude that our approach is efficient enough for our purposes of retrieving spatial and temporal information from the HDS articles.

The comparison and interpretation of precision across different IR systems is difficult, and thus a comparison of an IR system with a simple baseline approach is common. For example, this was illustrated in Derungs (2014: 83), who compared the results of their algorithm with an approach which randomly assigns referent locations to potential toponyms in case of toponym ambiguity. Such a comparison with a baseline approach was not made in this thesis but could be employed in future work.

[75] Hikr: http://www.hikr.org/ (accessed August 2016)

In this section, we have illustrated that the spatial and temporal information retrieval algorithms and the computation of spatio-temporal toponym relationships provide satisfying results in the creation of stable *spatialized network* visualizations. We could also demonstrate that the general network structure is only marginally affected by incorrectly retrieved spatial and temporal information from the HDS or incorrectly disambiguated spatial and temporal information by the GIR algorithm. In the next step, we illustrate the influence of the parameter choices on the *spatialized network* visualizations.

6.2 Assessment of parameters for the spatialized networks

In this section, we evaluate the *sensitivity* of the *spatialized networks* to the parameters chosen to compute the spatio-temporal relationships for this project. We decided to analyze the 19th century network of *Switzerland* for the same reasons as explained in *Section 6.1*. With this evaluation, we aim at assessing whether the *spatialized networks* and the characteristics of the networks we presented in *Subsection 5.2.1* are stable to parameter adaptations. One example of such a parameter is the 50% minimum rule for temporal references, as described in *Section 6.1*. We first present the parameters tested in this evaluation and subsequently discuss the evaluation results.

Within the sections titled *network parameters* in Figures 57-65, the parameters we chose for creating the spatio-temporal toponym relationships are listed. The definition of spatio-temporal relationships in this project and all parameters are detailed in *Subsection 4.2.1*. Figure 57 depicts the original *spatialized network* for the 19th century (see Figure 39), and thus all parameters listed in the *network parameters* section are equal to the definition in *Subsection 4.2.1*. In contrast, in Figures 58-65, one of the five parameters is adapted. The adapted parameters are highlighted in grey in the *network parameters* section.

The first parameter we adapted is the *toponym ranking method*. As introduced in *Subsection 2.1.3*, several methods exist which could potentially be used to rank terms (i.e., toponyms) in articles according to their relevance. We decided to incorporate the *Okapi BM25* method, as it is a standard method used in GIR to rank toponyms, and it incorporates *article length* in contrast to other common methods in GIR such as the *tf-idf*. We compared *Okapi BM25* to *tf-idf* in order to determine if this decision has an influence on network structure and characteristics (see Figure 58). The second parameter in Figures 57-65 is *POC and MUN excluded*. In *Subsection 4.1.1*, we introduced the exclusion of *places of citizenship* (i.e., POC) in *biographies*, and the membership of *municipalities* and *former municipalities* (i.e., MUN) to cantons and to districts in *geographical entities* article texts, as we are not interested in analyzing *places of citizenship* and the spatial hierarchies of cantons and districts in this project. We assessed the potential influence of excluding the POCs and MUNs from the networks, which is shown in Figure 59.

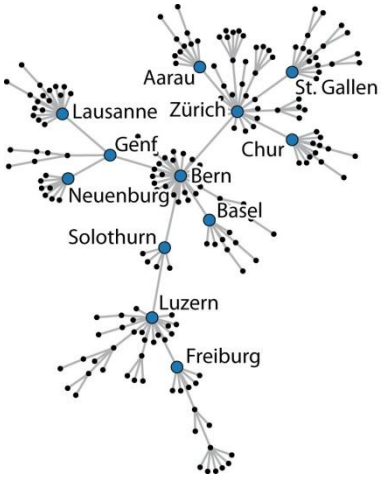
In addition to the *toponym ranking method* and the *exclusion of POCs and MUNs*, we tested the *minimum article number*, the *minimum temporal weight*, and the *minimum article rank* criteria (i.e., third, fourth and fifth parameter in Figures 57-65) because we are interested in how sensitive the structure of the *spatialized networks* and the network characteristics are to

these quantitative parameters we set for the computation of the spatio-temporal toponym relationship networks. For this reason, we adapted the criteria to be either more restrictive or less restrictive regarding the inclusion of articles in the *spatialized networks*, and analyzed how the network structure and characteristics might change. The justification for applying these parameters to compute the spatio-temporal relationships is detailed in *Subsection 4.2.1*.

As shown in *Subsection 4.2.1*, we defined that only toponym relationships occurring in at least three articles (i.e., *minimum article number* in Figures 57-65) are considered for the *spatialized networks*. We evaluated the effect of having no *minimum article number* criterion (i.e., less restrictive, see Figure 60), and the effect of a *minimum article number* which is twice as high as the criterion set for in this project (i.e., more restrictive, see Figure 61). The fourth parameter in Figures 57-65 is the *minimum temporal weight* as we only considered HDS articles for this project which have at least 50% of the temporal references to the studied century (i.e., 19th century for this evaluation). We evaluated the *sensitivity* of the *spatialized network* and the network characteristics to the reduction of this criterion to 33% (see Figure 62), and 0.1% (see Figure 63) for the computation of spatio-temporal toponym relationships. Finally, we assessed whether the *spatialized networks* and network characteristics change by defining a less restrictive *minimum article rank* (i.e., 0.0 in Figure 64) criterion and a more restrictive *minimum article rank* (i.e., 0.2 in Figure 65) criterion than the 0.1 criterion we have defined for the original networks. The 0.1 criterion says that only the 90% strongest spatio-temporal relationships per century are considered for computing the toponym relationship networks.

The number of *nodes* and *edges*, and the *average shortest path* (ASP) are displayed in the left column of the *network characteristics* sections in Figures 57-65. The *shortest path* represents the lowest number of edges between two nodes in a network. For example, the *shortest path* between *Luzern* and *St. Gallen* in Figure 57 contains four edges. The ASP is the average *shortest path length* for all possible pairs of nodes in a network. The more centralized a network is, the lower the ASP. Details and the motivation to use the ASP are presented in *Subsection 5.2.1*. In the right column of the *network characteristics* section, the influence of *biographies*, *thematic contributions*, *geographical entities*, and *families* to the *spatialized networks* is shown. In order to compute these parameters, the weighted toponym relationships (see Figure 27) were separated by article categories. For example, the more certain toponyms co-occur in *biographies* about the 19th century, the higher the percentage of *biographies* in the network *characteristics* section in Figures 57-65.

Network visualizations for the respective parameter configurations are depicted in the left half of Figures 57-65. Nodes are colored in black, and edges are displayed in grey. The 12 most central nodes in the reference network in Figure 57 are labeled in all networks. The evaluation particularly focuses on these specific nodes. In order to highlight their position in the network, they are shown in blue and depicted larger than the other nodes in the network. *Spatialized networks* were computed and visualized following the description in *Subsection 4.2.1*.



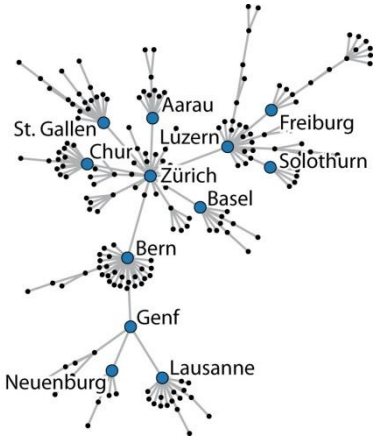
Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	3
Minimum temporal weight	50%
Minimum article rank	0.1

Network characteristics

Nodes	198	*Biographies*	92.4%
Edges	197	*Thematic contributions*	4.4%
Average Shortest Path	4.96	*Geographical entities*	2.1%
		Families	1.1%

Figure 57: *Okapi BM25* Switzerland reference network for the 19th century.



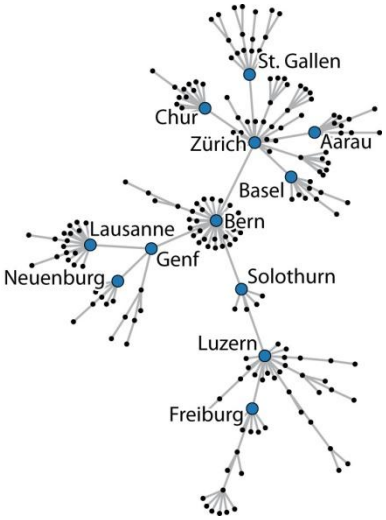
Network parameters

Toponym ranking method	tf-idf
POC and MUN excluded	yes
Minimum article number	3
Minimum temporal weight	50%
Minimum article rank	0.1

Network characteristics

Nodes	198	*Biographies*	92.6%
Edges	197	*Thematic contributions*	3.4%
Average Shortest Path	4.61	*Geographical entities*	2.2%
		Families	1.8%

Figure 58: *tf-idf* 19th century network of Switzerland.



Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	no
Minimum article number	3
Minimum temporal weight	50%
Minimum article rank	0.1

Network characteristics

Nodes	199	*Biographies*	92.1%
Edges	198	*Thematic contributions*	4.7%
Average Shortest Path	5.02	*Geographical entities*	2.2%
		Families	1.0%

Figure 59: *Okapi BM25* including place of citizenship and municipalities.

As a result of the applied layout algorithm in Figures 57-65, the positions of the nodes change across networks; therefore, absolute locations of nodes in the networks are not meaningful. However, we decided to always depict *Zürich* to the "northeast" (i.e., top right) of *Bern*, thus similar to the reference network in Figure 57. This also corresponds with the geographical arrangement, as *Zürich* is located northeast of *Bern* geographically. In the following, we first detail the reference network in Figure 57, then we describe the network structure in general, and finally we turn to the characteristics of the networks in Figures 58-65 and compare them to the reference network in Figure 57.

The reference network in Figure 57 contains 198 nodes and 197 edges. The *biographies* dominate the network, as they account for 92.4% of the weighted toponym relationships in the network, whereas other categories contribute 7.6% in total. *Zürich* and *Bern* are very centrally located in the middle of the network (i.e., *network hubs*) and are connected to many other nodes. The network has an ASP of 4.96. The theoretical maximum ASP (i.e., completely linear structure) for a network consisting of 198 nodes is 66.3, whereas the theoretical minimum ASP (i.e., completely centralized structure, all but one nodes are connected to one central node) is 2.0 (as described in *Subsection 5.2.1*). Therefore, the ASP of 4.96 indicates that the network is centralized, which is as expected due to the *pathfinder network scaling* we applied to the network that reduces complex and large networks to the structurally most relevant relationships and highlights hierarchical structures in the network (as described in *Subsection 4.2.1*). We further calculated a randomized network of 198 nodes, applying the *pathfinder network scaling*. The ASP of this randomized network is 11.8. Comparing this value to the ASP of 4.96 for the network in Figure 57 indicates that the reference network is more strongly centralized than a randomized *pathfinder network scaled* network, which meets our expectations of a spatial hierarchy of Swiss toponyms (e.g., central nodes such as *Zürich* which are directly connected to many less central nodes such as municipalities of the agglomeration of *Zürich*), being represented in the network. The theoretical maximum and minimum ASPs for all networks in Figures 57-65 are similar due to the similar number of nodes and edges for the networks and are therefore not discussed for each network separately. The same interpretation is valid for all networks in Figures 57-65: the networks are more centralized than a randomized network. The differences of the ASP in the networks in Figures 57-65 are only marginal (i.e., minimum of 4.61 in Figure 58, maximum of 5.17 in Figure 63) compared to the theoretically possible ASP range (i.e., difference from theoretical minimum to maximum ASP). This is an indicator of the stability of the network structure despite changing parameter settings.

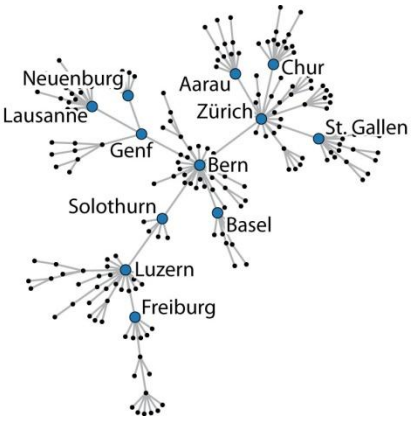
Applying different *toponym ranking methods* (i.e., *Okapi BM25* and *tf-idf*) or varying network parameters (e.g., *minimum article number*) has an influence on the connections in the network. For example, *Luzern* is directly connected to *Zürich* in Figure 58 instead of via *Solothurn* to *Bern*, as in the reference network in Figure 57. This effect is a result of the *pathfinder network scaling* algorithm we applied in order to depict only the structurally most representative relationships. *Luzern* and *Solothurn* are very strongly connected to each other and are thus directly connected in all networks. By investigating the connections of these two toponyms to other central nodes, we discovered that *Solothurn* is very

strongly connected to *Bern*, whereas *Luzern* is very strongly connected to *Zürich*. Depending on the strength of these two relationships, either the structure in Figure 57 (i.e., if *Solothurn-Bern* is stronger than *Luzern-Zürich*) or the structure in Figure 58 (i.e., if *Luzern-Zürich* is stronger than *Solothurn-Bern*) applies, which is an artifact of the *pathfinder network scaling* algorithm. The difference in strength of *Solothurn-Bern* and *Luzern-Zürich* is very small in all networks; therefore, the change in methods (i.e., *Okapi BM25* vs. *tf-idf*) or a slight variation in parameters has a large influence on the connections of these toponyms. A similar effect is noted for *Basel*. In four of the nine networks, including the reference network, *Basel* is directly connected to *Bern*. In the other five networks, *Basel* is connected to *Zürich*. This is due to an almost equal strength of relationship to *Bern* and *Zürich*, and thus the change of methods or varying network parameters influences the location and connections of *Basel*. Despite these instabilities, the network structure in Figures 57-65 is quite stable and the connections of central nodes and their connections to other nodes are consistent to a large degree.

We now turn to the single networks in Figures 58-65, and begin with Figure 58. The network characteristics in Figure 58 are similar to Figure 57. The contributions of *families*, *biographies*, and *geographical entities* are slightly higher compared to the reference network, whereas *thematic contributions* articles contribute less. The numbers of nodes and edges are equal. Therefore, applying *tf-idf* instead of *Okapi BM25* does not strongly influence the network structure and characteristics. In Figure 59, the network contains one more node and one more edge, which is attributable to *Klingnau* for the toponym relationship *Klingnau-Basel*. We applied the same procedure as discussed in *Section 6.1* to evaluate the relationship *Klingnau-Basel* and detected that it is only included because two out of three total *biographies* refer to *Klingnau* as a *place of citizenship* of people. However, we argued in *Subsection 4.1.1* that we wish to exclude *places of citizenship* from the computation of toponym relationships. Therefore, applying the algorithm to exclude *places of citizenship* from articles (see *Subsection 4.1.1*), thus excluding the relationship *Klingnau-Basel* from networks, as completed for the reference network (i.e., Figure 57), is reasonable.

In Figures 60 and 61, the influence of defining a *minimum article number* criterion different from three is illustrated. The contribution of the article categories is similar to the reference network in both cases, whereas the number of nodes and edges varies substantially. In Figure 60, a network is depicted which is based on at least one article per toponym relationship, causing all 203 toponyms to be considered. We evaluated the added nodes and toponym relationships, and determined that all added toponym relationships are relevant (following the evaluation procedure in *Section 6.1*). In Figure 61, a minimum of 6 articles has been applied; as a result, 19 nodes and edges were excluded, when compared to the reference network. We evaluated the toponym relationships of all 19 excluded nodes and determined that 14 of them are relevant, whereas 5 are irrelevant according to the evaluation procedure in *Section 6.1*. Therefore, applying 6 as *minimum article number* criterion would increase precision at the *network level* from 97% (see *Section 6.1*) to 99%, but the amount of nodes, and thus the recall, would decrease: if we assume a recall of 100% for the reference network in Figure 57, the recall

in Figure 61 would decrease to 90.4%, as only 179 out of 198 nodes are contained in the network. To summarize, the influence of adapting the *minimum article number* criterion is particularly high if the criterion is more restrictive (i.e., minimum of 6 articles in Figure 61). Although the *precision* increases, *recall* decreases by 10%, which is undesirable, as we wish to present a most complete toponym network to potential users. In contrast, Figure 60 (i.e., minimum of 1 article) illustrates a complete network, but for this network, five toponym relationships exist which are based only on a single article in which two toponyms co-occur. We decided that this is not representative enough for a toponym relationship being depicted in the *spatialized network*. Therefore, we selected three articles as a *minimum article number*.



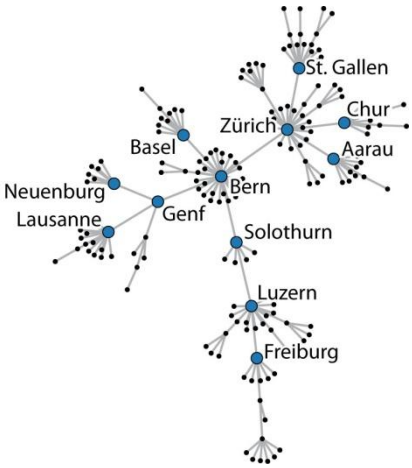
Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	1
Minimum temporal weight	50%
Minimum article rank	0.1

Network characteristics

Nodes	203	*Biographies*	92.2%
Edges	202	*Thematic contributions*	4.4%
Average Shortest Path	5.06	*Geographical entities*	2.3%
		Families	1.1%

Figure 60: *Okapi BM25* with at least one article.



Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	6
Minimum temporal weight	50%
Minimum article rank	0.1

Network characteristics

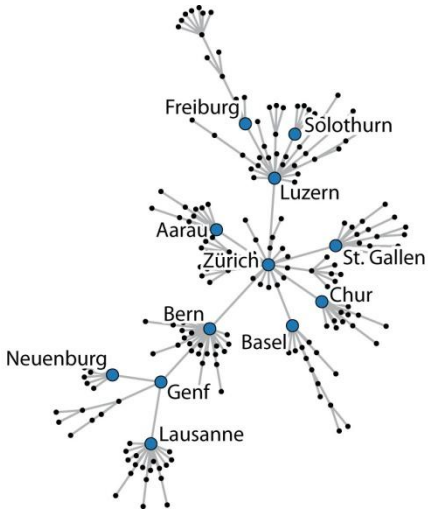
Nodes	179	*Biographies*	92.4%
Edges	178	*Thematic contributions*	4.9%
Average Shortest Path	4.75	*Geographical entities*	1.7%
		Families	1.0%

Figure 61: *Okapi BM25* with at least six articles.

In Figures 62 and 63, we decreased the limit for including articles to 33% and 0.1% temporal references of the 19th century in each article. Both networks incorporate all 203 nodes. The contribution of articles differs substantially. For example, for the network in Figure 63, *biographies* contribute 17% less than to the reference network, whereas *thematic contributions*, *families*, and *geographical entities* contribute 84%, 218%, and 481 % more, respectively, than to the reference network in Figure 57. The strong contribution of *thematic contributions*, and particularly of *families* and *geographical entities*

articles by decreasing the temporal weight criterion in Figures 62 and 63, is primarily due to the fact that articles in these categories often cover a wide time range and thus are disregarded with the 50% limit, though included with lower limits. For example, HDS articles regarding the cantons of Switzerland (i.e., *geographical entities* article category) cover many centuries and thus score low for single centuries. The *Canton of Zurich* article, for example, contains 23.4% temporal references to the 19th century. The remaining 76.6% are divided into other centuries. Similarly, the *thematic contributions* article about *agriculture* covers the entire history of agriculture in Switzerland and thus only 28.3% of the temporal references are about the 19th century. Both articles are disregarded with the 50% temporal references limit, but included with the lower limits. In contrast, *biographies* cover shorter time periods and are therefore better attributable to single centuries.

Our choice for the 50% criterion for temporal references was based primarily on our interest in depicting networks which are based on content (i.e., articles) regarding an analyzed time period (i.e., the 19th century in this evaluation). Therefore, only articles which mainly (i.e., 50% of the temporal references) refer to a century were considered.



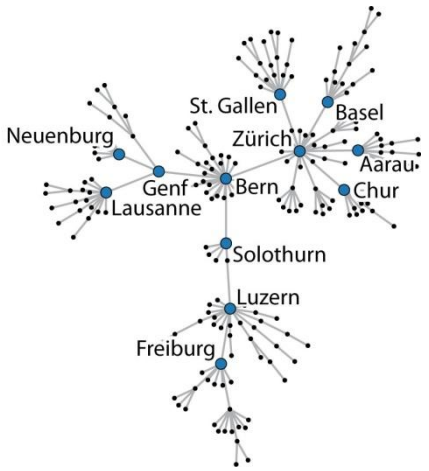
Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	3
Minimum temporal weight	33%
Minimum article rank	0.1

Network characteristics

Nodes	203	*Biographies*	85.6%
Edges	202	*Thematic contributions*	7.0%
Average Shortest Path	4.76	*Geographical entities*	5.3%
		Families	2.1%

Figure 62: *Okapi BM25* with temporal limit 33%.



Network parameters

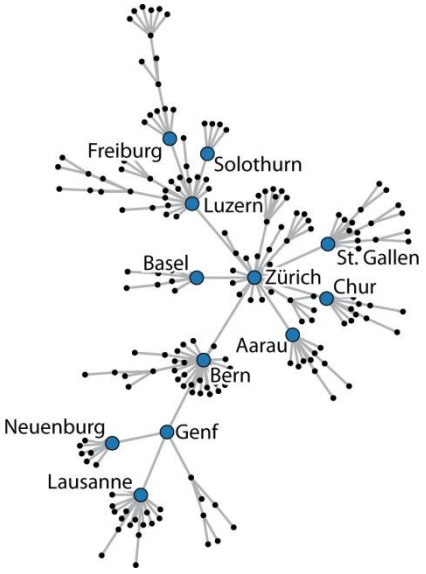
Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	3
Minimum temporal weight	0.1%
Minimum article rank	0.1

Network characteristics

Nodes	203	*Biographies*	76.2%
Edges	202	*Thematic contributions*	8.1%
Average Shortest Path	5.17	*Geographical entities*	12.2%
		Families	3.5%

Figure 63: *Okapi BM25* with temporal limit 0.1%.

In order to calculate the networks in Figures 64 and 65, we applied different *minimum article rank* criteria. In Figure 64, all articles were considered. In Figure 65, only the 80% strongest articles regarding their contribution to spatio-temporal networks were investigated. The network characteristics are similar to the reference network. However, in Figure 64, *Biasca* is additionally included and connected to *Bellinzona*. We evaluated this additional toponym relationship and judged it as relevant. *Bellinzona*, which is not labeled in Figure 64, is directly connected to *Genf*. The variation of a *minimum article rank* causes a higher contribution of *thematic contributions* articles and a lower contribution of *biographies* in Figure 64 when compared to the reference network. In contrast, a lower contribution of *thematic contributions* articles and a higher contribution of *biographies* articles are shown in Figure 65. This is due to the substantially higher article length and lower number of toponyms per 100 words of *thematic contributions* articles when compared *to biographies* (see *Subsection 5.1.1*). Therefore, many *thematic contributions* articles have a low toponym relationship strength and are thus excluded with a more restrictive *minimum article rank* criterion, as *Okapi BM25* considers article length (i.e., the longer the articles, the lower the *Okapi BM25*). The percentage of contribution of *geographical entities* and *families* articles in Figures 64 and 65 is similar to the reference network.



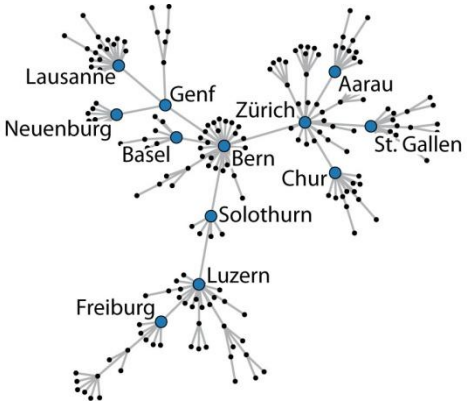
Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	3
Minimum temporal weight	50%
Minimum article rank	0

Network characteristics

Nodes	199	*Biographies*	91.7%
Edges	198	*Thematic contributions*	5.1%
Average Shortest Path	4.64	*Geographical entities*	2.1%
		Families	1.1%

Figure 64: *Okapi BM25* with no minimum article rank.



Network parameters

Toponym ranking method	Okapi BM25
POC and MUN excluded	yes
Minimum article number	3
Minimum temporal weight	50%
Minimum article rank	0.2

Network characteristics

Nodes	198	*Biographies*	93.6%
Edges	197	*Thematic contributions*	3.2%
Average Shortest Path	4.96	*Geographical entities*	2.1%
		Families	1.1%

Figure 65: *Okapi BM25* with 20% minimum article rank.

Our choice to select 90% as a *minimum article rank* criterion was particularly based on the aim to exclude very long articles with many toponyms (e.g., the *thematic contributions* article about the *industrialization*) which would otherwise connect a massive number of toponyms with extremely weak spatio-temporal toponym relationships.

In this evaluation, we focused on the most central nodes of the *spatialized network* in the 19th century and conclude that the network structure is quite stable regarding the tested parameters, with the exception for *Luzern*, *Solothurn*, and *Basel*. These toponyms are sensitive to parameter changes, as previously discussed. The decision for many parameters was based on factors such as the project aim of only incorporating articles which are attributable to a single century (i.e., *minimum temporal weight*), and the aim to exclude very long articles with many toponyms (i.e., *minimum article rank*) but we could not detect a major influence (i.e., *sensitivity*) of our choice for these parameters on the general network structure and network characteristics.

In the following sections, we investigate the thematic information retrieved from the HDS articles. We first illustrate the evaluation procedure applied in order to determine an optimal number of topics to be considered for *topic modeling* in *Section 6.3*. Then, the chosen model is compared against a manual classification of the HDS *thematic contributions* articles in *Section 6.4*.

6.3 How many topics for topic modeling?

In this section, we describe and evaluate the approach we chose to select a meaningful number of topics for the *topic modeling* (TM) out of the 3,067 *thematic contribution* articles. As introduced in *Subsection 4.1.3*, we employed MALLET in order to compute the TM. In the output of MALLET, the *log likelihood/token* (LLT) is produced. The higher the LLT, the better the model fit, and thus the accuracy of the TM (Griffiths and Steyvers, 2004, Wallach et al., 2009b: 1106-07). We computed LLT for 2 to 60 topics since, in the early stages of the HDS project, the people responsible for defining themes to be covered in the HDS had planned to include approximately 22 different themes in the *thematic contributions* article category (personal communication, HDS, 2016a). We assumed 20 to 30 topics to be optimal for the TM, and assessed a wider range in the evaluation to verify our assumptions. As LLT values slightly vary from run to run, we computed five LLTs for each number of topics, averaged these values, and visualized the result in a line chart, which is depicted in Figure 66. In this figure, the blue line represents the exact LLT based on the number of topics, and the black line is a polynomial trend line which highlights the general trend of the LLT in the chart. The polynomial trend line was selected because it performs particularly well for approximating fluctuating data, as is the case for the data represented in Figure 66. The trend line in Figure 66 is typical for model fit evaluations for TM (e.g., Griffiths and Steyvers, 2004, Barbieri et al., 2013): the LLT is high for only two topics and then decreases for a greater number of topics. For less than ten topics, the LLT is low, which implies that the TM solution does not accurately represent the thematic content of the articles. The LLT increases with higher numbers of topics, and the trend line reaches a

peak between 20 and 30 topics. This implies that these TM solutions fit very well, and thus these models represent the thematic content of the articles very well. We expected this peak at approximately 20 to 30 topics, as previously mentioned. The LLT decreases with higher numbers of topics, except for a small peak between 50 and 60 topics.

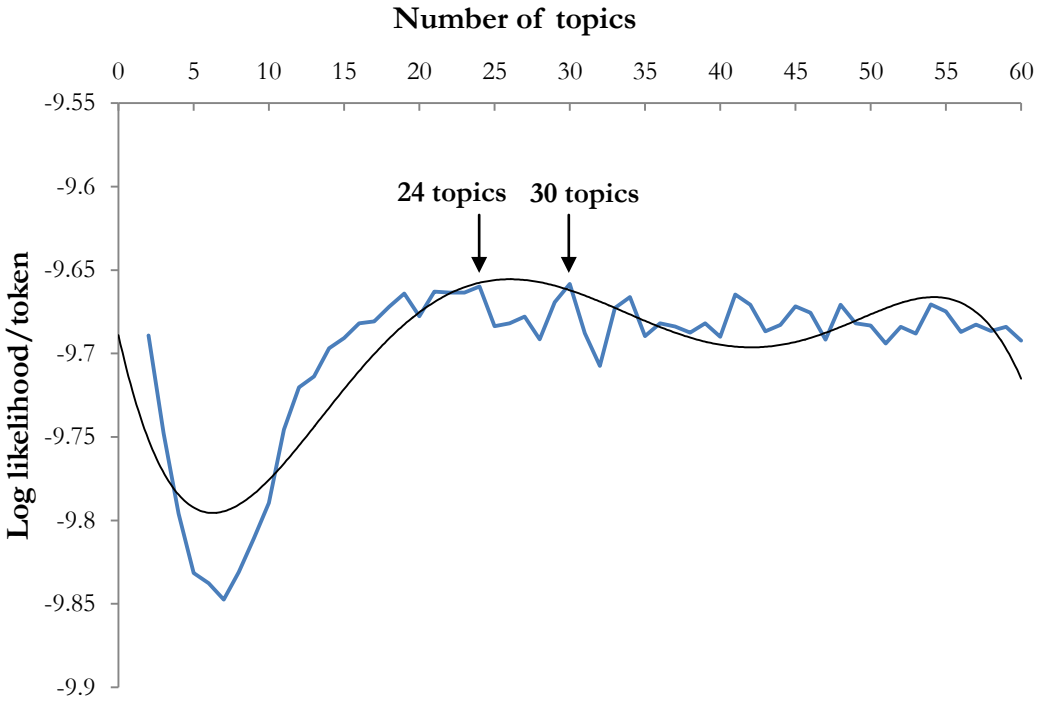


Figure 66: *Log likelihood/token* values for different numbers of topics.

We identified the highest LLTs for the 24 (i.e., -9.660) and 30 (i.e., -9.658) topics solutions. We decided to compute *thematic landscapes* for both of these solutions in order to compare them qualitatively and visually to one another, and to decide on one of them to be employed in our project. This decision is described in *Subsection 4.1.3*, as we followed the suggestions of Chang et al. (2009) and Mimno et al. (2011) to not only use quantitative measures in order to find an appropriate number of topics for the TM (e.g., LLT), but also include qualitative methods (i.e., human judgments).

The retrieval of thematic information from the HDS articles and the computation of the *thematic landscape* for the 30 topics solution are described in *Subsections 4.1.3* and *4.2.2*, respectively. The 24 topics solution was computed in the same way as the 30 topics solution by applying the same parameters. However, the automatic thematic clustering of the *thematic contributions* articles according to Blondel et al. (2008) resulted in 22 themes for the 24 topics solution, compared to 28 themes for the 30 topics solution. The *thematic landscapes* for both the 22 themes and 28 themes, are visualized in Figures 67 and 68.

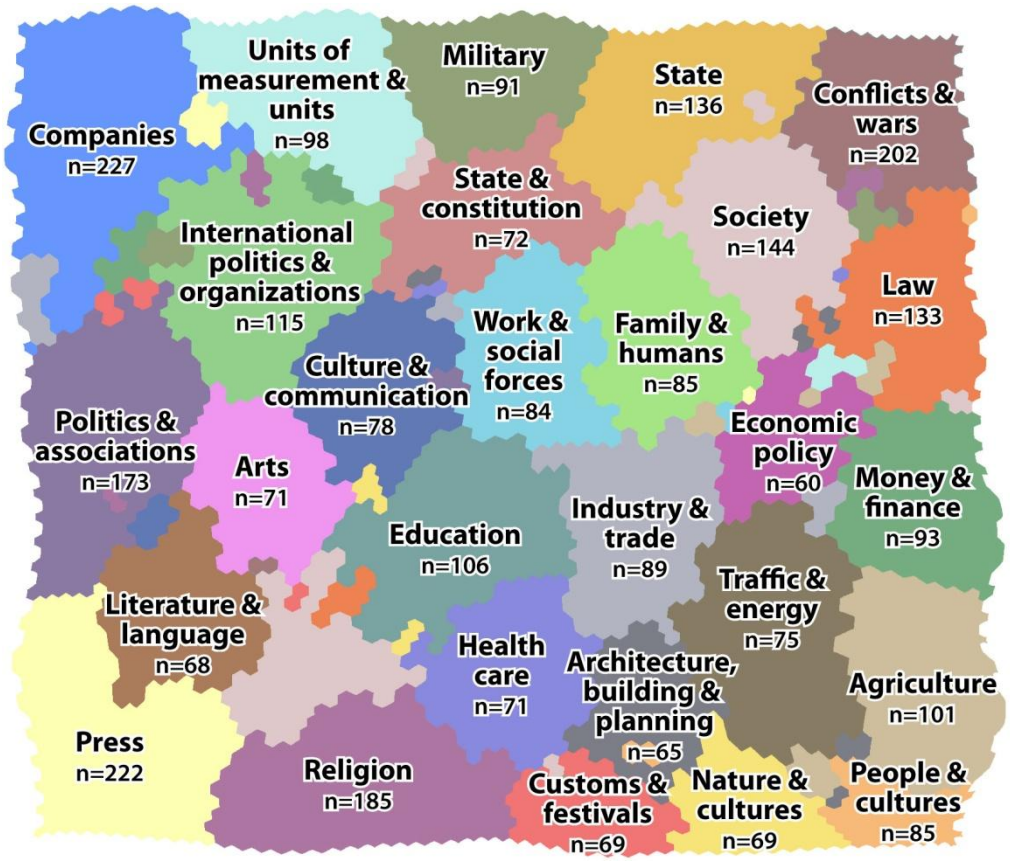


Figure 67: *Thematic landscape* with 28 themes.

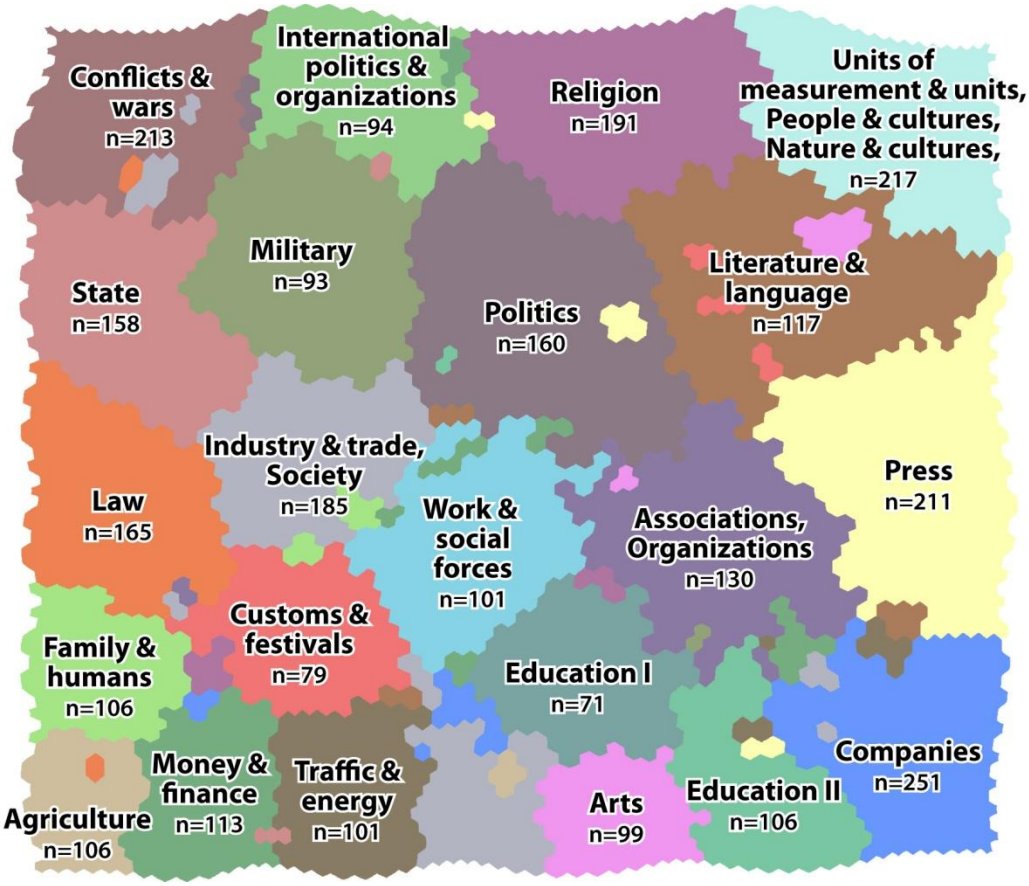


Figure 68: *Thematic landscape* with 22 themes.

We analyzed the membership of articles to the themes in the landscapes in Figures 67 and 68. We used the themes in Figure 67 as a reference and analyzed to which themes the articles were assigned to in Figure 68. Labels were defined by assessing which theme in Figure 67 contributed most to each of the themes in Figure 68. We added the labels of themes which represent at least 50 articles in a theme in Figure 68 to the label. For example, *Industry & trade* contributes the most articles to the grey theme in Figure 68. In addition, 50 articles classified as *Society* in Figure 67 are present in this grey theme in Figure 68; therefore, the label contains both theme labels (i.e., *Industry & trade, Society* in Figure 68). The number of articles (= *n*) per theme is displayed below the labels. The color hues for the themes in Figure 68 are identical to those in Figure 67.

The landscape in Figure 68 contains 23 out of 28 themes, which are displayed in Figure 67 (e.g., *Religion*, *Military*). *Units of measurement & units*, *People & cultures*, and *Nature & cultures* articles, which are single themes in Figure 67, form one large theme in Figure 68. *Industry & trade* and *Society* form another combined theme. *Education* is divided into two themes in Figure 68 and *Politics & associations* is split into *Politics* and *Associations, Organizations*. The labels for *Politics* and *Associations, Organizations* were defined by the author of this thesis by reading the titles of articles in these themes and specifying a descriptive term. Five themes from Figure 67 are missing in Figure 68: *Economic policy* articles in Figure 67 are primarily assigned to the *Money & finance* theme in Figure 68, while most *State & constitution* articles are part of the *State* theme, and the major part of the *Health care* articles are assigned to the *Family & humans* and *Education II* themes. *Architecture, building & planning* articles are found in the *Traffic & energy* as well as in the *Arts* theme in Figure 68, and most articles of the *Culture & communication* theme in Figure 67 are divided into *Press*, *Arts*, and *Customs & festivals* in Figure 68.

The size of many themes in the two landscapes is comparable. However, for some themes, differences occur which are due to, among other reasons, the merging of themes from Figure 67 to Figure 68. For example, in Figure 68, 49 additional articles are part of the *Literature & language* theme compared to 68 articles in Figure 67. Most of these 49 articles (i.e., 30 articles) were classified as *Press* in Figure 67. We revealed that articles of these two themes (i.e., *Literature & language* and *Press*) are very thematically similar, which is supported by the arrangement of the themes in the *thematic landscape* (i.e., they are neighboring themes in both figures). We further identified that most of the 30 articles which were classified as *Press* in Figure 67, and as *Literature & language* in Figure 68, are located in the border region of these themes in both figures. For example, HDS articles in this border region include *printing* (= *Buchdruck*) and *publishing* (= *Verlage*). The content of these articles is related to both themes and thus their location in the border region is reasonable. The reason why these articles belong to the *Press* theme in Figure 67 and to the *Literature & language* theme in Figure 68 might be primarily due to an adapted thematic focus of the *Press* theme in Figure 68, when compared to Figure 67. The reason for this is that the *Culture & communication* theme in Figure 67 does not exist anymore in Figure 68, and the articles which were classified as *Culture & communication* in Figure 67 were assigned to different themes in Figure 68, such as the *Press* theme. Therefore, the *Press* theme in Figure 68 contains several articles

(i.e., 18 articles) which were classified as *Culture & communication* in Figure 67. As a consequence, the thematic focus of the *Press* articles was moved slightly towards the topic *Culture & communication*, and thus the articles in the border region of *Press* and *Literature & language* fit better into the *Literature & language* theme than into *Press*.

Based on these findings, we found the more detailed representation of the HDS *thematic contributions* articles, with 28 themes (see Figure 67), as being more reasonable than the 22 themes solution (see Figure 68) for the purpose of our project. This is because we wish to present information seekers with an overview of themes present in the HDS. However, with the 22 themes solution, several themes covered in the HDS are not represented (e.g., *Culture & communication*, *economic policy*, *health care*), or become clustered to combined themes. One example of such a combined theme in Figure 68 is *Units of measurement & units*, *People & cultures*, *Nature & cultures*. We analyzed articles of this theme and studied the location of the articles within the blue region in the top right corner of Figure 68. We read several HDS articles in this region, and discovered that the articles related to *Units of measurement & units* are thematically different from articles in the *People & cultures* and *Nature & cultures* theme, and thus combining them is not reasonable. This finding is supported by the fact that the *Units of measurement & units* articles are located in a separate region from the other two themes within the *Units of measurement & units*, *People & cultures*, *Nature & cultures* theme in Figure 68 (i.e., the larger the distance between articles in the *thematic landscape*, the less thematically similar they are). This is not shown in Figure 68, though we discovered this by considering the *detail view* (i.e., individual articles view) of the *thematic landscape*. These findings support our choice for the 28 themes solution. We conclude that choosing only 22 themes might be not sufficient to cover the thematic diversity of the HDS *thematic contributions* articles.

In this section, we have illustrated the evaluation process of finding an optimal topic number as an input for the *topic modeling*. Our quantitative approach revealed that the 24 and the 30 topics solutions are optimal, and thus we depicted both solutions in *thematic landscapes* in order to compare them visually to each another. From the thematic clustering of the articles, we obtained 22 themes for the 24 topics solution and 28 themes for the 30 topics solution. The constitution and visual arrangement of these themes in the *self-organizing maps* in Figures 67 and 68 are similar. However, the qualitative (i.e., visual) assessment of the two solutions revealed that the 30 topics/28 themes solution fits better for the purpose of this project, as previously detailed. In the following section, the selected 28 themes solution is further evaluated by comparing it to a manually annotated classification of the HDS.

6.4 Comparing HDS article clustering methods

In this section, we illustrate an evaluation of two different clustering approaches for the HDS *thematic contributions* article category. We compare the *Blondel communities* that we computed for the *thematic contribution* articles (i.e., the 28 *themes*) to a manually annotated classification scheme of *thematic contributions* articles. This manually annotated

classification scheme was created as a part of the *New HDS* project (see *Section 3.3*). Thematic information was assigned to all 3,067 *thematic contributions* articles by human annotators (i.e., HDS employees), who classified the articles manually into a three-level hierarchical system. On the highest level, seven categories exist, and each article belongs to at least one of these categories. For each article, there is also information on a second (= 93 categories), and a third hierarchical level (= 668 categories). For this comparison, we only considered the highest level categories, as this is the only level which has similar cluster sizes to the *themes*, as we further detail in the proceeding text. The prototypical data set of the manually annotated HDS articles was provided courtesy of the HDS on April 29, 2016, and at the time of this thesis being written the data has not yet been published by the HDS.

In the course of this evaluation, we employed statistical measures and compared the two different clustering methods to one another by means of visual inspection.

Evaluation approach

In Table 21, the two compared clustering methods are illustrated for the four articles studied in detail in *Chapter 5 – Results*. We also included the article *Agrarpolitik* (= *agricultural policy*). We incorporated this article to explain a special case, which is detailed later. For each article, a *theme* and an *HDS class* exists, as shown in Table 21.

Table 21: Themes and HDS classes.

Article	Theme	HDS class (manual classification)
Crossair	Companies	Economy
Militärorganisationen (MO)	Military	State, power, law, politics
Journal du Jura	Press	Culture, arts, science, religion, mentalities
Matin, Le	Press	Culture, arts, science, religion, mentalities
Agrarpolitik	Economic policy	Economy & State, power, law, politics

The *theme* of an article refers to the thematic clustering applied to the *article-topic matrix*, as detailed in *Subsections 4.2.2* and *5.2.2*. The *HDS class* refers to the manual classification of the articles by the HDS team. As previously described, at the highest hierarchical level, the HDS contains seven categories. Each article was assigned manually to one or several of these categories by HDS editorial office employees. For example, the article *Crossair* in Table 21 is categorized as *Economy*, the article *Militärorganisationen (MO)* is categorized as *State, power, law, politics*. Different HDS categories are visualized with a corresponding color in Table 21. The article *Agrarpolitik* differs from the other articles in Table 21, as it was assigned two categories by the HDS: *Economy* and *State, power, law, politics*. We considered such combinations of categories as one *HDS class* for our comparative analysis. Following this procedure, it is possible to conduct an *article-to-article comparison* of the *HDS classes* and the *themes* because each article is part of only one *HDS class* and one *theme* at a time.

In total, 96 of these *HDS classes* exist, each comprising one *HDS category* or several of the seven *HDS categories*. However, for this evaluation, we only considered *HDS classes*

which contain at least 60 articles. This choice was based on the fact that the smallest *theme* (= *Economic policy*) contains 60 articles, and thus we aimed to assess comparable cluster sizes. We obtained 12 *HDS classes* with more than 60 articles. These 12 *HDS classes* are listed in Table 24. For each class, the categories are listed and, similar to Table 21, are depicted using different colors. As only a few articles were categorized by the HDS as *historical science and identity formation* and *environment, space, settlement*, these two *HDS categories* are missing from Table 24.

We first look at Table 21 and compare the automatically generated *themes* and the manually created *HDS classes* for the five listed articles. For example, both the *Crossair* and the *Agrarpolitik* articles were assigned *themes* which are related to *economy* (i.e., *Companies*, *Economic policy*). This semantically matches with the HDS class *Economy*. The article *Agrarpolitik* was assigned to *Economy* and additionally the HDS class *State, power, law, politics*. This is reasonable, as agriculture is an *economic* sector, and the article reports on how agriculture is organized in Switzerland by the *state* and other (political) organizations (Baumann and Moser, 2012). The article about *Militärorganisationen (MO)* was also classified as *State, power, law, politics*, and automatically assigned the theme *Military*, which is reasonable because it describes the organization of the *Swiss army*, the administration of the *Swiss army*, and military service in general, including military training and education (Senn, 2010). The articles *Journal du Jura* and *Matin, Le* are very similar in content (as demonstrated in *Subsection 5.1.3*) because both describe daily newspapers in the French-speaking region of Switzerland and were thus automatically assigned the theme *Press*. Similarly, these two articles were assigned the same *HDS class* (i.e., *Culture, arts, science, religion, mentalities*). Therefore, for the five articles in Table 21, the thematic clustering of the articles is comparable regarding the *themes*/*HDS classes* which were assigned. In the next step, we compared the *themes* and *HDS classes* statistically to one another.

In Table 23, the *themes* and *HDS classes* are compared to one another in a matrix. In the rows, the *themes* are listed and the *HDS classes* are shown in the columns. The abbreviations used for the *HDS classes* in the column headings are specified in Table 24. The numbers in each cell of Table 23 reveal how many articles of a *theme* were classified as the *HDS class* specified in the respective column. For example, 201 articles are part of the theme *Companies* and the same 201 articles were manually classified as *Economy* by the HDS (i.e., HDS class *h* in Table 23). The row for the theme *Companies* is framed by a black box in Table 23 as we refer several times to this theme in the proceeding text. One article of the 201 *Companies* articles is *Crossair* (see Table 21).

The highest frequency in each row of Table 23 is highlighted in bold. For example, in the *Companies* row, 201 is highlighted which implies that most articles of the *Companies* theme were manually classified by the HDS as *Economy* (i.e., HDS class *h*). In the rightmost column, the number of articles for each *theme* that are covered by the 12 *HDS classes* in total is displayed, and, in brackets, the percentage of all articles in the respective *theme* is provided. For example, the theme *Companies* contains 227 articles. In total, 225 articles of these 227 articles were assigned by the HDS to the 12 *HDS classes* in Table 23, and thus the value in brackets is 99% (225/227 = 0.991). The table is sorted

by this percentage from largest to smallest. In the row at the bottom of Table 23, the total number of articles per *HDS class* is shown. *Economy* (i.e., column *h*) represents the largest *HDS class* with 548 articles, which is highlighted in bold. In total, the 12 *HDS classes* cover 2,182 of 3,067 *thematic contributions*, which results in 71% coverage.

We further statistically analyzed how well the *themes* and *HDS classes* correspond to each other by applying the *hypergeometric test* from Kos and Psenicka (2000). The null hypothesis is that the membership of articles to the *themes* and *HDS classes* only corresponds by chance (Kos and Psenicka, 2000: 859). Rejecting this hypothesis supports the assumption that the two clusters are similar (i.e., correspond). The calculation for the *hypergeometric test* is depicted in Formula 1. The variables used are explained in Table 22.

Formula 1: *Hypergeometric test* (Kos and Psenicka, 2000: 859).

$$p = 1 - \sum_{j=0}^{NTH-1} \frac{\binom{NH}{j}\binom{N-NH}{NT-j}}{\binom{N}{NT}}$$

Table 22: Defining *hypergeometric test* variables (Kos and Psenicka, 2000: 860).

Variable	Definition
N	Total number of *thematic contributions* in the HDS (= 3,067)
NH	Number of articles in the *HDS class*
NT	Number of articles in the *theme*
NTH	Number of articles in the *theme* that match with the *HDS class*

In statistical terms, Formula 1 computes the probability of drawing *NTH* or more successes with a sample size of *NT* within the finite population of size *N* containing *NH* possible successes (Kos and Psenicka, 2000: 859). Each draw is either a success or a failure, and draws are not replaced (Kos and Psenicka, 2000: 859). In other words, the lower the resulting *p* value, the higher the probability that the articles in the two clusters do not only correspond by chance. For example, the probability that 201 out of 225 articles of the *Companies* theme (see Table 23) are assigned to the same HDS class (i.e., class *h*) by chance is very low. Therefore, the *Companies* theme and the HDS class *h* correspond stronger than expected for the case that each of the 225 *Companies* articles would have been randomly assigned to one of the 12 *HDS classes*. This implies that the HDS class *h* is significantly overrepresented in the *Companies* theme. We applied Formula 1 to all cells in Table 23 and colored them in blue if the null hypothesis was rejected. Darker shaded cells in Table 23 indicate that the null hypothesis was rejected at a lower significance level (i.e., the probability computed with Formula 1 is lower) compared to lighter shaded cells. This implies that the chance that articles in the respective *theme* and the *HDS class* correspond by chance is very low. White cells in Table 23 indicate that there is no statistically significant correspondence between the *theme* and the *HDS class*.

Table 23: A comparison of themes to HDS article classes.

Theme \ HDS class	a	b	c	d	e	f	g	h	i	j	k	l	Total
Companies	3				1			**201**	3	10	6	1	*225 (99%)*
Press	**194**	1	3		1	2		9	3			1	*214 (96%)*
Units of measurement & units			1					**85**		4	1	1	*92 (94%)*
Money & finance								**60**	1	20	1		*82 (88%)*
State	3	4	**87**	15		2	4	1		1			*117 (86%)*
Family & humans	2		1		**24**	14	10				20	2	*73 (86%)*
Health care	2		1		**16**	**16**	2	2	1	1	11	7	*59 (83%)*
Military		3	**50**	3	2		8			7		1	*74 (81%)*
Customs & festivals	4		1		4	**17**	1	4	4		2	13	*50 (72%)*
Politics & associations	4	7	**42**	6	15	8	10	9	1	10	6	4	*122 (71%)*
Work & social forces			1		3			**26**	3	3	21	2	*59 (70%)*
Conflicts & wars		**101**	16	16	1		6						*140 (69%)*
Religion	**97**		1			16	1		4		1	7	*127 (69%)*
Industry & trade	1							**39**	5	2	8	6	*61 (69%)*
Traffic & energy								**36**	5	4	2	4	*51 (68%)*
Education	**37**				2	11		2	8		1	8	*69 (65%)*
State & constitution		1	**19**	4			3	3	1	14			*45 (63%)*
International politics & organizations	5	12	12	**16**	7	1	7	1	2	6	2		*71 (62%)*
Agriculture			1					**46**	1	6	8		*62 (61%)*
Literature & language	**23**		2		1	7		3	2		1	2	*41 (60%)*
Society	14	2	13	4	11	4	**22**		3	1	1	8	*83 (58%)*
Culture & communication	**13**		1		2	5		1	11	2	3	6	*44 (56%)*
Nature & cultures		**22**			6	2		8					*38 (55%)*
Law	2		**30**	10	3		13			7	1		*66 (50%)*
People & cultures	1	**26**	1	12			1	1					*42 (49%)*
Economic policy					4			7		**9**	7		*27 (45%)*
Arts	**13**		1			2		1	9			4	*30 (42%)*
Architecture, building & planning					2			3	**7**		2	4	*18 (28%)*
Total	*418*	*179*	*284*	*86*	*105*	*107*	*88*	***548***	*74*	*107*	*105*	*81*	*2182*

Significance level of group consistency

.001 .01 .05 Not significant

Table 24: HDS classes and categories.

HDS class	HDS category/categories
a	Culture, arts, science, religion, mentalities
b	Chronological approach
c	State, power, law, politics
d	Chronological approach & state, power, law, politics
e	Society, population, way of living
f	Culture, arts, science, religion, mentalities & society, population, way of living
g	State, power, law, politics & society, population, way of living
h	Economy
i	Economy & culture, arts, science, religion, mentalities
j	Economy & state, power, law, politics
k	Economy & society, population, way of living
l	Economy & culture, arts, science, religion, mentalities & society, population, way of living

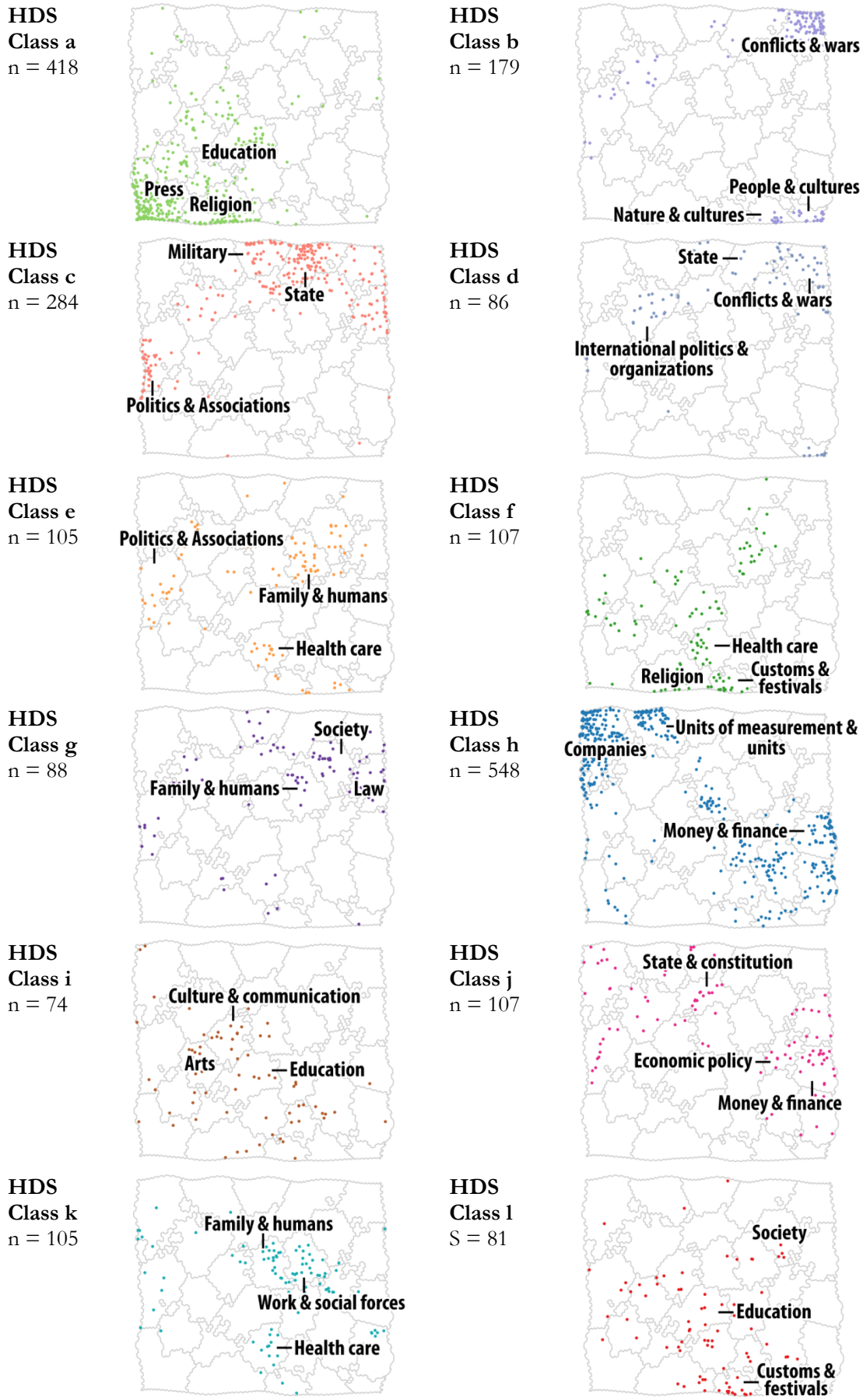


Figure 69: HDS classes and themes in *thematic landscape*.

In the next step, we visually compare the *themes* to the *HDS classes*, as illustrated in Figure 69. We decided to conduct this comparison because we wished to test our expectation that articles of the same *HDS class* cluster in the same region of the *thematic landscape*. We expected this due to the *distance-similarity metaphor*, which applies to *self-organizing maps* (Fabrikant et al., 2006). As such, similar articles (i.e., articles from the same *HDS class*) should be placed in close proximity in the *thematic landscape*. The articles of the *HDS classes* are visualized as differently colored dots on top of the *thematic landscape* in Figure 69. The grey polygons in the *thematic landscape* represent the *themes*. In each of the 12 maps, the three *themes* that are most frequent in a *HDS class*, and significantly correspond with the *HDS class* at least at the 0.05 level (according to the results in Table 23), are labeled. To the left of the maps, the *HDS class* abbreviations as well as the size of the classes (= *n*) are displayed.

Evaluation results

We first consider the coverage in Table 23 which is displayed in the rightmost column. Coverage ranges from 99% for the theme *Companies* (i.e., almost all *Companies* articles are contained by the 12 HDS classes) to 28% for the theme *Architecture, building & planning*. Despite these large differences in coverage, for each theme (i.e., each row) there is at least one statistically overrepresented HDS class (i.e., blue cells in Table 23). This implies that each theme thematically corresponds with at least one HDS class. For example, in Table 23, the cell for the theme *Companies* and the HDS class *h* (i.e., *Economy*) is colored in blue. This implies that the HDS class *Economy* is statistically overrepresented (i.e., a strong thematic correspondence) in the theme *Companies*, as previously mentioned. In the following paragraphs, we discuss, for each HDS class, how they are represented in the 28 themes in order to uncover how themes and HDS classes correspond. Furthermore, we extend our findings by analyzing the article distribution of the HDS classes in the *thematic landscape* (see Figure 69).

The HDS class *Economy* is overrepresented in seven themes in Table 23: *Companies* (201 articles), *Units of measurement & units* (85 articles), *Money & finance* (60 articles), *Agriculture* (46 articles), *Industry & trade* (39 articles), *Traffic & energy* (36 articles), and *Work & social forces* (26 articles). These themes are all thematically related to economy, and thus the correspondence with the HDS class *Economy* is not surprising. The fact that *Economy* contains 548 articles in total (i.e., 25% out of all articles in Table 23) highlights the prevalence of this HDS class in Table 23. This prevalence is also visible in Figure 69, as the HDS class *h* articles cover a large area in the upper left part, the center, to the right of the center, and at the bottom right part of the *thematic landscape*. HDS classes that are built up by *Economy* and another HDS category (i.e., HDS classes *i*, *j*, *k*, *l*, see Table 24) contain 367 articles (i.e., 17% out of all articles in Table 23). In contrast to HDS class *h*, articles of these classes are less concentrated in regions of the *thematic landscape*. This is reasonable as these classes relate to themes other than *economy* as well, and are thus less clearly attributable to single themes and regions in the *thematic landscape*.

The HDS class *Culture, arts, science, religion, mentalities* (i.e., HDS class *a*) contains a large number of articles, and is overrepresented in four themes from Table 23: *Press* (194 articles), *Religion* (97 articles), *Education* (37 articles), and *Literature & language* (23 articles). These themes are all thematically related to *Culture, arts, science, religion, mentalities*. Therefore, the overrepresentation of this HDS class in the aforementioned themes is to be expected. *Culture, arts, science, religion, mentalities* is slightly less prevalent than *Economy* in Table 23 and contains 418 articles (i.e., 19% out of all articles in Table 23). HDS class *a* articles are present in the region at the bottom left corner and between the bottom left corner and the center of the *thematic landscape* in Figure 69. HDS classes which are a combination of *Culture, arts, science, religion, mentalities* and another HDS category (i.e., HDS classes *f*, *i*, *l*, see Table 24) contain 262 articles (i.e., 12% out of all articles in Table 23). The articles of these classes are less clustered in a specific region of the *thematic landscape* than HDS class *a* in Figure 69, but are generally close to the *Culture, arts, science, religion, mentalities* class.

The HDS class *State, power, law, politics* (i.e., HDS class *c*) contains substantially less articles (i.e., 284 articles) than *Economy* and *Culture, arts, science, religion, mentalities*, but is still overrepresented in five themes in Table 23: *State* (87 articles), *Military* (50 articles), *Politics & associations* (42 articles), *Law* (30 articles), and *State & constitution* (19 articles). Articles of the HDS class *c* are located in the left center, top center, and top right corner of the *thematic landscape* in Figure 69. For HDS classes that are a combination of *State, power, law, politics* and another HDS category (i.e., HDS classes *d*, *g*, and *j*), we find a similar pattern, as found for the HDS classes discussed previously (i.e., *Economy* and *Culture, arts, science, religion, mentalities*). Articles of such combined HDS classes are more dispersed over the *thematic landscape* and less clustered in a region compared to the HDS class, which only consists of one HDS category (i.e., HDS class *c*).

The HDS class *chronological approach* (i.e., HDS class *b*) contains fewer articles (i.e., 179 articles) than the HDS classes previously discussed and is overrepresented in the *Conflicts & wars* (101 articles), *People & cultures* (26 articles), *Nature & cultures* (22 articles), and *International politics & organizations* (12 articles) themes in Table 23. HDS articles about these themes typically report on a chronological sequence, and thus the overrepresentation of the HDS class *chronological approach* is reasonable. For example, articles regarding wars (i.e., theme *Conflicts & wars*) typically describe the war in chronological order. Articles typically begin describing the time before the war, then report on important dates and events during the war, and finally report on the consequences following war. Articles of HDS class *b* are located at the top right and bottom right corner in Figure 69. HDS class *d* is a combination of the *chronological approach* class and the *state, power, law, politics*, and is located close to the HDS class *b* in Figure 69, particularly in the top right corner of the *thematic landscape*.

The HDS class *Society, population, way of living* (i.e., HDS class *e*) contains the lowest number of articles (i.e., 105 articles) for HDS classes that consist of one HDS category (i.e., HDS classes *a*, *b*, *c*, *e*, *h*) in Table 23. It is overrepresented in the *Family & humans* (24 articles), *Health care* (16 articles), *Politics & associations* (15 articles), *Society* (11 articles), and *Nature & cultures* (6 articles) themes. In Figure 69, HDS class *e* is dispersed over the

thematic landscape and is not strongly clustered in any specific regions of the *thematic landscape*, which is similar to HDS classes that are a combination of the *Society, population, way of living* and another HDS category (i.e., HDS classes *f*, *g*, *k*, and *l*). Therefore, articles that were classified by the HDS as *Society, population, way of living* are related to many of the 28 themes, but without being overrepresented with many articles in a specific theme, no strong clustering of HDS class *e* articles is displayed in Figure 69.

The evaluation reveals a strong correspondence of *themes* and *HDS classes*, which is illustrated by many cells being colored in blue in Table 23. This implies that the membership of articles to the two clusters of articles (i.e., *themes*, *HDS classes*) is not caused by chance because many articles which are part of the same theme (e.g., *Companies*) are also part of the same HDS class (e.g., *Economy*). Therefore, the 28 automatically computed themes in this project correspond well with the manually annotated classification scheme (i.e., the *HDS classes*) of the articles. Visual inspection of the HDS class articles in the *thematic landscape* in Figure 69 revealed that, except for *Society, population, way of living* HDS classes, which are based on one HDS category only (i.e., HDS classes *a*, *b*, *c*, *h*), are concentrated in specific regions of the *thematic landscape*. This is reasonable, as the *self-organizing map* algorithm allocates articles which are similar in close proximity in the *self-organizing map* (i.e., the *thematic landscape*), as detailed in *Subsection 4.2.2*. Articles of HDS classes that consist of at least two HDS categories (e.g., articles of the HDS class *j*: *Economy* and *State, power, law, politics*) are less clearly attributable to individual themes and specific regions within the *thematic landscape* in Figure 69. This is reasonable, as these articles are related to different themes in the *thematic landscape* and are thereby less clearly attributable to single themes and regions.

In summary, at the beginning of this chapter we stated that we tested the *quality* and the *sensitivity* of our results to the methods applied and parameters chosen in this project. We first demonstrated that the spatio-temporal information retrieval and computation of spatio-temporal relationships works well in the context of this project, and potential errors in the disambiguation process would only have a small influence on network structure. Furthermore, we illustrated that the network structure is stable and only marginally influenced by our parameter choices. Next, we evaluated the optimal topic number for the TM in a combined quantitative and qualitative/visual approach. This procedure helped us to decide on the 30 topics/28 themes solution, as it fits the underlying data best. In the final section of this chapter, we compared the chosen solution with 28 themes to a manually annotated classification scheme of the HDS. We discovered a high correspondence of *themes* and *HDS classes*, which indicates that the automatically generated themes correspond well with the human classification scheme of the articles. In the next section, these findings are incorporated in a discussion of the results of this thesis.

7 Discussion

In the humanities, a strong interest to analyze large online text archives with computational methods has initiated the formation of the *digital humanities* discipline, which has attracted the interest of a large interdisciplinary community of researchers in recent years (Kaplan, 2015). Within this discipline, several sub-fields such as the *GeoHumanities* and *spatial humanities* have evolved to tackle research questions at the nexus of the humanities and geography. We illustrated an approach situated in this emerging and interdisciplinary field between geography and the humanities in this thesis, and proposed to take advantage of typical GIScience methods to address information-seeking needs in the humanities. In this chapter, we discuss the results of applying our approach in this interdisciplinary field between geography and the humanities in the context of this thesis' research questions. We place the results in the context of related research discussed in Chapter 2 of this thesis, and highlight new insights we gained as well as limitations we uncovered. We start this chapter by discussing the results of retrieving spatial, temporal, and thematic information from a typical humanities digital text archive.

7.1 Revisiting the research questions

In *Chapter 1 – Introduction* of this thesis, we introduced three research goals for this project. The first goal is to automatically extract spatio-temporal and thematic information from large digital text archives in the humanities so that hidden structures and relationships can be uncovered in the data. This is challenging, as many text archives in the humanities are unstructured or semi-structured, and thus spatial, temporal, and thematic information needs to be identified and retrieved first. We considered automatic text processing algorithms to be relevant for the identification and retrieval of spatio-temporal and thematic information because manual extraction is very time-consuming for large text archives. Therefore, we introduce the following research question.

Research Question 1

How can information about space, time, and theme be automatically retrieved from unstructured and semi-structured text archives in the humanities so that hidden structures and relationships can be uncovered in the data?

We chose the *Historical Dictionary of Switzerland* (HDS) as a case study for this project because it is a typical and representative semi-structured digital text archive in the humanities containing a wealth of spatial, temporal, and thematic data, as illustrated in *Section 5.1*. These data have not been retrieved from the HDS or systematically analyzed and depicted, and only limited querying options are available in the current online version of the HDS, as shown in *Chapter 3 – Data*. We studied the German version of the HDS in this project, which is described in *Section 3.2*.

To retrieve spatial, temporal, and thematic information from the HDS, and thus to answer *Research Question 1*, we identified methods suggested by the *geographic information retrieval* (GIR) community to be relevant, as these methods have been optimized for automatically retrieving spatio-temporal and thematic information from large unstructured and semi-structured text archives. In addition, the methods provide answers to deal with ambiguity in geographic data (e.g., does *London* refer to *London, UK* or *London, Ontario, Canada*?) which was introduced in *Subsection 2.1.1*.

We first applied a GIR algorithm developed by Derungs and Purves (2014) to retrieve spatial information from the HDS articles. We chose this algorithm as it is optimized for retrieving spatial information from unstructured German texts. We retrieved 322,179 toponyms in total and detected that 97.9% of the 36,188 articles include at least one toponym. This substantially exceeds the often stated assertion in GIScience that 80% of all information has a reference to space or geography (e.g., MacEachren and Kraak, 2001: 3). Hahmann and Burghardt (2013: 1186) studied spatial information on *Wikipedia* and found that 57% of the articles are geospatially referenced, whereas Adams and Gahegan (2016) report that 75% of the *Wikipedia* articles contain at least one toponym (populated and unpopulated places), and 54% of the articles include at least one toponym referring to a populated place. These numbers are significantly lower than the 97.9% of articles we found to contain spatial information in the HDS. This is not surprising and indicates that typical text archives about history, such as the HDS, contain substantially more articles related to space than text archives which are thematically more diverse, such as *Wikipedia*. This is because text documents about history have a very strong spatial component as they report on specific sites (i.e., places) where human actions take place (Ayers, 2010: 1-3). We detected that the high density of spatial information in the HDS is particularly attributable to *populated places* (i.e., 82% of all toponyms in the HDS). This is also not surprising, as most human actions occur where people live.

We then applied the *HeidelTime* tool, which we selected because it provides an automatic method to retrieve temporal information from unstructured and semi-structured text

documents (Strötgen and Gertz, 2013). As a result, we were able to retrieve 499,258 temporal references from the HDS. We determined that 99.5% of all HDS articles contain at least one temporal reference. Adams and Gahegan (2016) retrieved temporal information from *Wikipedia* articles and report that 93.2% of all *Wikipedia* articles contain at least one temporal reference. The higher number of articles related to time in the HDS, compared to the thematically more diverse *Wikipedia* data archive, is also not surprising, as *history happens over time* (Gregory, 2010: 58), and the HDS covers a wide period of time, ranging from the *Paleolithic* until *today*. Recent time periods are covered much more frequently in the HDS than less recent time periods. This is due to the adapted strategy of the HDS to have 40% of the HDS content to be about the 19th and 20th century (Morosoli, 2000: 10). We found that even 59% percent of the temporal references cover the 19th and the 20th century. This implies that these centuries are substantially overrepresented regarding the initial plans of the HDS.

We then set out to evaluate the *quality* of our spatio-temporal information retrieval approach, as detailed in *Section 6.1*. Therefore, we employed a typical *system-oriented* GIR evaluation approach (see *Subsection 2.1.4*) and decided to evaluate the *precision* (e.g., Manning et al., 2009b). We obtained a *precision* of 83% for *articles*. This implies that from 83% of the HDS articles considered to create the *spatialized network visualizations*, the spatial and the temporal information was correctly retrieved and disambiguated by the GIR algorithm. We compared our *precision* to previous work in GIR, as illustrated in *Section 6.1*. We chose the work of Derungs (2014) and Palacio et al. (2015) because they applied a very similar GIR approach to retrieve spatial information from unstructured text documents in German. Derungs (2014: 86) reported a *precision* of 82% of their own algorithm, and Palacio et al. (2015) a precision of 70-90%. Our *precision* is thus very similar. This highlights that existing spatial and temporal information retrieval techniques work well with the HDS text archive.

We further retrieved thematic information from the HDS articles by applying *topic modeling* (TM), and found a TM solution with 30 topics to thematically best represent the 3,067 *thematic contributions* articles in the HDS. To evaluate the number of topics which best fits the data, we applied a combined quantitative and qualitative approach, as suggested by Chang et al. (2009) and Mimno et al. (2011). A TM *article-topic matrix* was used to cluster articles by thematic content, applying the *community detection algorithm* developed by Blondel et al. (2008). As a result, we obtained 28 themes. The clustering of the articles was graphically and statistically compared to a manually annotated classification of the HDS *thematic contributions* articles, as shown in *Section 6.4*. The two clustering solutions were found to be consistent, which indicates that the considered approach to automatically retrieve and cluster thematic data in the HDS complies with human judgments of similarity. Thus, to include 30 topics and 28 themes, respectively, in this project seems reasonable for thematically clustering the HDS text archive.

In research communities situated at the nexus of geography and the humanities, such as the *digital and spatial humanities* (see *Subsection 2.4.2*), GIR techniques to automatically retrieve spatial, temporal, and thematic information from large digital text archives have

been promoted by several researchers in recent years (e.g., Berzak et al., 2011, Clifford et al., 2016, Donaldson et al., 2016). Similar to the approach shown in this thesis, the use of gazetteers to retrieve toponyms from large digital text archives is a common strategy in the *digital and spatial humanities* (e.g., Gregory and Hardie, 2011, Simon et al., 2015). Existing gazetteers and automatic GIR systems have been optimized for automatically retrieving spatial information from humanities text documents, and particularly from historical text document collections (e.g., Grover et al., 2010, Southall et al., 2011, Alex et al., 2015, Gregory et al., 2016). Automatically clustering the thematic content of large text archives in the humanities has been investigated by scholars in the *digital and spatial humanities*, as well, as shown by Jockers (2013), who applied a *topic modeling* approach similar to the one presented in this thesis. For example, a combined approach to retrieve spatio-temporal and thematic information from digital text archives in the humanities has been illustrated by Hinrichs et al. (2015). Information about space, time, and theme was retrieved from a large historical text document collection to study commodity trading in the 19th century by applying a semi-automatic information retrieval approach (see *Subsection 2.4.2*). However, no combined spatio-temporal and thematic information retrieval approach has been suggested by the interdisciplinary *digital and spatial humanities* research community that allows us to automatically identify and extract all three dimensions (i.e., space, time, and theme) from large unstructured or semi-structured digital text archives in the humanities to date. Such an approach would facilitate detecting spatio-temporal and thematic structures and relationships in unstructured and semi-structured text documents in the humanities. The lack of this type of approach was indicated as a research gap in *Section 2.5*. We promote the approach presented in this thesis to the *digital and spatial humanities* to bridge this gap, and particularly highlight the potential use of the retrieved information about space, time, and theme to uncover and visualize hidden structures and interconnections in the text documents for further data exploration, which is discussed by answering *Research Question 2*, presented next.

To summarize, we have illustrated that the HDS contains a great deal of spatial, temporal, and thematic information. This information has not yet been systematically retrieved from the HDS, as illustrated in *Chapter 3 – Data* of this thesis. In response to *Research Question 1*, we find that spatial, temporal, and thematic information can indeed be automatically retrieved from an unstructured or semi-structured humanities text archive by applying well-established *geographic information retrieval* methods. Visualizing the retrieved information to support the information seeker in the humanities to gain new insights into spatio-temporal and thematic structures and relationships in humanities text documents is the second goal of this research project. Therefore, we introduce the following research question.

Research Question 2

How can we spatialize uncovered spatio-temporal and thematic structures and interconnections extracted from unstructured and semi-structured text archives in the humanities?

Moretti (2005: 1-2) suggested visualizing such structures and interconnections by applying a *distant reading* approach, which is illustrated by the following notion we introduced in *Chapter 1 – Introduction*.

> *"(…) literature, the old territory (more or less), unlike the drift towards other discourses so typical of recent years. But within that old territory, a new object of study: instead of concrete, individual works, a trio of artificial constructs—graphs, maps and trees—in which the reality of the text undergoes a process of deliberate reduction and abstraction. 'Distant reading', I have once called this type of approach; where distance is however not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.*
> *(…)*
> *And so, while recent literary theory was turning for inspiration towards French and German metaphysics, I kept thinking that there was actually much more to be learned from the natural and the social sciences."*
>
> Moretti (2005: 1-2)

Moretti (2005) highlights *graphs*, *maps*, and *trees* as potential visual displays to depict structures and interconnections in large text data archives in the humanities. We followed Moretti's (2005) *distant reading* approach and visualized spatio-temporal and thematic structures and relationships uncovered in the HDS in *graphs* (i.e., *network visualizations*) and *maps* (i.e., *self-organizing map*) which is further detailed next.

We decided to depict the spatio-temporal information retrieved from the HDS text archive in network visualizations, as we were interested in the visualization of interconnections and structures inherent in the spatio-temporal data that we automatically retrieved from the HDS. Visualizations should cognitively support people to understand such interconnections and structures in large text data, compared to reading and extracting such information from masses of text (i.e., *close reading*). For this reason, we applied the *spatialization framework*, theoretically developed by Fabrikant and Skupin (2005). This framework suggests a systematic transformation and reorganization to depict multidimensional data with lower dimensional visualizations, using *spatial metaphors*. The network spatialization concept is based on the *first law of cognitive geography*, which states that spatializations are experienced by people as if they were geographic spaces (i.e., people *believe* closer things are more similar than distant things) (Montello et al., 2003). Therefore, the *distance-similarity metaphor* is applied, and items which are semantically more similar are placed closer to one another in the network space compared to items which are semantically dissimilar (Fabrikant et al., 2004).

We followed a structured approach to transform and visualize the spatio-temporal HDS data. To compute spatio-temporal networks, we applied the *Okapi BM25* (Manning et al., 2009c) algorithm to measure the relevance of toponyms in each article. We then extended it with a temporal weighting factor that describes the percentage of temporal references per century in each article. Next, we analyzed co-occurrences of toponyms in HDS articles about the same century, following Salvini (2012) and Hecht and Raubal (2008), who analyzed co-occurrences of spatial and thematic information in *Wikipedia* articles. Our combined spatio-temporal relationship score was used as an input to create network spatializations. The network layout (i.e., *graph embedder* (GEM) in combination with the *pathfinder network scaling*) we applied in this project highlights hierarchical structures in the toponym networks (see Figures 38-41), because only structurally most relevant relationships are depicted (Börner et al., 2003: 201-03). A similar procedure to highlight hierarchical structures of a toponym network was applied by Salvini (2012), who analyzed the toponym relationships at a global level, which is different from our project, as we investigated toponym networks at a national and a cantonal level (i.e., *Switzerland* and *Canton of Zurich*). Salvini (2012) observed that *New York*, *London*, and *Paris* are the most central *network hubs* by analyzing the *Wikipedia* hyperlink structure. In contrast to Salvini (2012), we analyzed toponym networks over time. This revealed, for example, that *Bern* and *Zürich* were most central in the 19th and 20th century, whereas in the 18th century, *Basel* and *Luzern* were most central and thus placed in the middle of the network. We illustrated that target users of our project (i.e., historians) were able to create such findings by interacting with the *spatialized network interface* (see *Subsection 5.3.3*). Thus, they were able to gain new insights into the dynamic organization of toponyms in networks about Switzerland over time. Therefore, we conclude that target users in the humanities might benefit by the spatialization method presented in this thesis because the approach supports uncovering spatio-temporal structures and interconnections in humanities text data archives.

Salvini (2012: 178) pointed to a limitation of the network layout approach we applied: the *pathfinder network scaling* algorithm visually highlights only the structurally most important relationships in a toponym network. This has two major implications: first, toponyms that are strongly connected to one another might be placed relatively distant from one another in the network visualization, because the layout is only optimized for the structurally most important relationships in the network. Therefore, such toponyms are interpreted by humans to be only weakly related due to the *distance-similarity metaphor*, which operates in network spatializations (Fabrikant et al., 2004). Second, for many toponyms, only the strongest relationship is displayed in the network (see Figures 38-41), and all other relationships are discarded. We applied two methods to compensate for these effects: first, we employed the *community detection algorithm* by Blondel et al. (2008), which delineates toponym clusters. A toponym cluster consists of densely related toponyms within a cluster, and weak relationships to toponyms outside a cluster (see *Subsection 4.2.1*). Toponyms that are part of the same community were assigned the same *color hue* in the network visualization. Due to the strength of the visual variable *color hue* (Bertin, 1967) for designating categories or classes (MacEachren, 1995), humans are supported in interpreting toponyms that are distant from one another in the

network, but part of the same community, as being strongly related to one another. Second, we calculated for each toponym the three strongest relationships to other toponyms in the network. These relationships are displayed in the interactive *spatialized network interface* (see Figure 47). Therefore, a target user is provided with more details about the toponym relationship network compared to if only the strongest relationships would be depicted. Salvini (2012) incorporated the first method (i.e., computing toponym clusters and coloring them differently) as well, whereas the second method (i.e., visually highlighting the three strongest toponym relationships interactively) is an extension of Salvini's (2012) approach.

As we depicted network spatializations of *Switzerland* and the *Canton of Zurich* in different centuries, we had to decide whether to preserve the *mental map layout* or not, which was discussed and evaluated by the *dynamic network visualization* community (e.g., Archambault and Purchase, 2012). The preservation of the *mental map layout* implies that the overall shape of the network is kept over time, and thus only as few nodes as possible are moved as little as possible, as introduced in *Subsection 2.2.3* (Archambault et al., 2011). Empirical evaluations in literature show contradicting results regarding the effect of preserving the *mental map layout* on the performance of users (i.e., error rate and response time) with dynamic network visualizations. For example, to test performance, users were asked to compare the centrality of nodes in different time periods (e.g., Archambault et al., 2011) or to memorize and remember the evolution of a dynamic graph sequence over time (e.g., Archambault and Purchase, 2012). Purchase and Samra (2008), Saffrey and Purchase (2008), Archambault et al. (2011), and Archambault and Purchase (2012) reported that *mental map preservation* has no influence on the performance of users. Archambault and Purchase (2013), however, reported a significantly better performance of users (i.e., fewer errors, faster response time) if the map layout is preserved. We did not conduct user studies to test whether target users perform better with or without *mental map layout preservation* in this thesis, but decided not to keep the *mental map layout* in different centuries because we wished to highlight differences in the hierarchical structure and the position of toponyms in different centuries and thus to optimize the layout in each time slice, as suggested by Branke (2001). The results of the second *think aloud study* (see *Subsection 5.3.3*) indicate that optimizing the layout of each time slice helped participants to detect particular characteristics of single time slices, such as identifying the most central nodes in the *spatialized network* or detecting changes of the network structure for Switzerland over time. For example, participants identified that the 18th century network is much more linear compared to the 19th and 20th centuries, and *Basel* and *Pruntrut* are located in the middle of the network (see *Section 5.3.3*). In contrast, *Zürich* and *Bern* are located in the middle of the network in the 19th and 20th centuries. These findings by participants thus support that not preserving the *mental map layout* is a reasonable strategy if particular characteristics of the network structure should be highlighted in time series. Thus, the results support the findings reported in Branke (2001).

To spatialize the thematic information retrieved from the HDS text archive, we employed the *self-organizing map* (SOM) technique. SOMs are a common *clustering* and

data reduction technique to depict bimodal relational data sets (e.g., an *article-topic matrix*) in a spatialized display (Skupin and Agarwal, 2008). We chose SOMs because they perform particularly well with large input data, as introduced in *Subsection 2.2.3*. Furthermore, we expected to support the information-seeking process of target users because the *distance-similarity metaphor* in SOMs (i.e., articles in close proximity are similar) should be intuitively understood by humans. This is because the *first law of cognitive geography* applies in this region-display spatialization, according to Fabrikant et al. (2006). As the similarity of neighboring neurons in a SOM is not equally distributed over a SOM, we applied the *cartogram* technique to distort the size of the neurons. Neurons that are similar to neighboring neurons are shrunk; neurons that are dissimilar to neighboring neurons are enlarged. Therefore, HDS articles in neighboring neurons that are similar (i.e., neighboring neurons are small) are placed closer to one another in contrast to articles in dissimilar neurons that are pushed apart from one another in the SOM cartogram. As a consequence, the SOM representation is optimized regarding the *first law of cognitive geography* (Fabrikant et al., 2006), because dissimilarity of neurons correlates better with distances of neurons in SOM cartograms than in traditional SOMs. This procedure was introduced by Salvini (2012) and Bruggmann et al. (2013).

In the SOM cartogram, themes were created by applying the *community detection algorithm* by Blondel et al. (2008) to the thematic information (i.e., *article-topic matrix*) retrieved from the HDS articles. As shown in *Subsection 5.2.2*, we created two hierarchical levels of the SOM: an *overview SOM*, showing all 28 themes as colored regions in the SOM, and a *detail view*, showing single HDS articles colored by the theme they were automatically assigned to (see Figure 42). This method followed the recommendation by Fabrikant et al. (2006) to depict thematic regions in the SOM in different colors to reinforce the similarity judgments of viewers. Depicting a SOM output on different hierarchical levels to analyze the thematic content of digital text data has already been suggested by Skupin (2002), Skupin and de Jongh (2005), and for spatio-temporal data by Andrienko et al. (2010a), as presented in *Subsection 2.2.3*. As a difference to previous work, we developed two hierarchical levels of SOMs to illustrate the thematic content of a typical semi-structured digital text archive in the humanities (i.e., different research contexts). We optimized the design of the two hierarchical levels of the SOM based on our goal to incorporate both levels into a user interface for interactive exploration. We provided target users with coupled *distant* and *close reading* options, as suggested by Jockers (2013), and as further detailed by answering *Research Question 3*.

By analyzing the two hierarchical levels of the SOM, we identified that semantically similar themes are located in similar regions of the SOM, which meets our expectations (see *Subsection 5.2.2*). For example, in the top right corner of the *thematic landscape* (see Figure 42), several themes to *state and society* are clustered. Thematically ambiguous themes are placed in the middle of the *thematic landscape*, which is a pattern that has already been detected by Salvini (2012) and in own previous work (Bruggmann et al., 2013), and is due to the thematic similarity of these themes to many other themes in the landscape. The interpretation of the SOM and thus the understanding of the *distance-similarity metaphor* (Fabrikant et al., 2006) by target users was investigated in

the second *think aloud study* (see *Subsection 5.3.3*). Users interpreted neighboring regions and articles in the SOM as thematically more similar than distant regions and articles. This result was expected because we optimized the SOM according to the *first law of cognitive geography* (Fabrikant et al., 2006) by applying the SOM cartogram technique and defined regions of thematically similar articles in the SOM, as previously described (see also *Subsection 5.3.3*).

The approach we presented in this thesis extends previous work regarding the empirical evaluation of *spatializations* (see *Section 2.2*). The *distance-similarity metaphor* in spatializations was empirically evaluated for network spatializations (Fabrikant et al., 2004) and region-display spatializations (i.e., SOMs) (Fabrikant et al., 2006) in previous work. For both spatialization types, the results of the empirical evaluations illustrated that the *distance-similarity metaphor* operates. However, these evaluations only incorporated the perceptual aspect of interpreting the *distance-similarity metaphor*, whereas semantics of data items were disregarded. This implies that the evaluated spatialized displays only included illustrative examples, and data items in the spatializations had no semantic meaning (i.e., attributes). Our approach extends this work, as we incorporated semantics in the spatialized displays (i.e., data items are specific documents with semantic content). For example, the points depicted in the SOM represent HDS articles and were assigned attributes (i.e., information about their thematic content). We were able to illustrate that the *distance-similarity metaphor* in spatializations operates with a *real* application in the humanities (i.e., text documents about history), and we can show that the *spatialization framework* can be used for depicting spatio-temporal and thematic structures and interconnections found in humanities text documents. The results of the evaluations with our target users will be further discussed by answering *Research Question 3*.

The spatialization approach presented in this thesis further contributes to previous attempts to spatialize information retrieved from text documents in the *digital and spatial humanities*. Using spatialized displays to depict structures and relationships found in text documents is not new. For example, several scholars have employed network spatializations to depict relationships between people (i.e., social networks) extracted from various document collections in the humanities (e.g., Ciula et al., 2008, Bingenheimer et al., 2011, Tóth, 2013), or highlighted relationships between texts of a corpus using network visualizations (e.g., Weingart and Jorgensen, 2013, Reiter et al., 2014). When it comes to spatial information retrieved from text documents, the use of maps, such as thematic (e.g., Murchú and Lawless, 2014) or density maps (e.g., Gregory and Hardie, 2011), is the most common visualization technique to date. Places are often visualized using graduated circles that are placed on a map (e.g., Hinrichs et al., 2015). To highlight spatial and spatio-temporal relationships, such as travel routes (e.g., Evans and Jasnow, 2014) or co-occurrences of place names in text documents, lines are superimposed on maps (e.g., Barker et al., 2010). However, the use of network spatializations to depict spatial and spatio-temporal information and relationships retrieved from unstructured or semi-structured digital text archives is mostly absent in the *digital and spatial humanities*. Thus, we attempt to bridge this gap with the approach presented in this thesis. Regarding the spatialization of thematic information, we found

several research projects that applied the SOM technique to text data in research contexts not related to the humanities. The spatialization of social media data (e.g., Steiger et al., 2016) or conference abstracts (e.g., Skupin and de Jongh, 2005) are two examples. For humanities-related disciplines, we found other spatialization techniques than SOMs applied to unstructured and semi-structured text archives, such as the *multidimensional scaling technique* (e.g., Mimno, 2012). However, we could not find any evidence in the *digital and spatial humanities* literature that show the use of SOMs to spatialize (thematic) information, automatically retrieved from unstructured or semi-structured digital text archives. Our spatialization approach thus fills a research gap in the *digital and spatial humanities* regarding the visualization of spatio-temporal and thematic structures and interconnections in the humanities. The potential of our spatialization approach to support information seekers in the humanities to gain relevant insights into space, time, and theme in text archives in the humanities has been demonstrated by the empirical evaluations of the two interactive web interfaces with respective target users. This will be discussed further in answering *Research Question 3*.

To summarize, we presented the *spatialization framework* as an answer to *Research Question 2*. We decided to depict spatio-temporal relationships in *network spatializations*, and thematic structures and interconnections in a *self-organizing map*. By answering *Research Question 2*, we followed Moretti's (2005: 1-2) data abstraction and reduction path: creating *graphs*, *maps*, and *trees* to allow *distant reading*. Similarly to Moretti's (2005) *maps*, *graphs*, and *trees*, we suggested *networks* (i.e., *graphs*) and *self-organizing maps* (i.e., *maps*) for *distant reading*. As Jockers (2013) suggested, we link *distant* reading with *close reading* concepts to provide information seekers an overview and details-on-demand access to digital text archives in the humanities, which is the third goal of this research project. Next, we answer and discuss the third and last research question.

Research Question 3

How can we make spatialized information about space, time, and theme from unstructured and semi-structured text archives available to information seekers in the humanities to support sense-making and the generation of new insights about these text archives?

We considered *geovisual analytics* (geoVA) relevant to answer *Research Question 3* because it provides a systematic framework to incorporate spatio-temporal and thematic data in exploratory web interfaces, with the aim to support the gain of new insights about latent spatio-temporal and thematic structures and interconnections buried in large (text) data archives (e.g., Andrienko et al., 2010b). In addition, we found geoVA relevant because it suggests systematic user-centered interface design and evaluation approaches to learn about needs and requirements of target users and to evaluate *utility* and *usability* of interactive user interfaces dealing with space, time, and theme (e.g., Roth et al., 2015).

In our project, we followed particularly the approach presented by Roth et al. (2015) because they applied a typical iterative user-centered interface design and evaluation

approach in a context related to this project (i.e., visualizing spatio-temporal and thematic data in an exploratory web interface). In previous work about user-centered interface design, the early involvement of target users in the interface design and evaluation process has been recommended to optimize interfaces according to information-seeking needs by target users (e.g., Lewis and Rieman, 1993). Therefore, we first organized a *focus group meeting* that helped us to assess requirements of our target users and to obtain feedback regarding our initial ideas of the planned spatio-temporal and thematic web interfaces. The results of the *focus group* were promising, as participants confirmed a strong interest in interactively exploring text data in the humanities from a spatio-temporal and thematic point of view. Additionally, *focus group* participants suggested new functionalities such as the combination of network with map views in interactive user interfaces. Furthermore, they convinced us to incorporate thematic information in exploratory interfaces. Therefore, we extended the interface concept with *self-organizing maps*. These findings were beneficial for us, as they helped us to adapt our initial plans for the interactive user interfaces. We believe that one reason for obtaining such detailed feedback on our interface ideas was that we used paper mockups, which encouraged the creativity of target users. This is because paper mockups provide target users with the impression that the interface concept is not yet complete, as reported in previous works (e.g., Wong, 1992, Landay and Myers, 2001). Getting to know the target users and their spatio-temporal and thematic information needs during the *focus group meeting* supported us in subsequent conceptual development of the interface ideas. Our findings highlight the usefulness and potential benefits of involving target users at a very early stage of the user-centered interface design and evaluation process, which is supported, for example, by Lewis and Rieman (1993) and Rubin and Chisnell (2008: 17).

In the next step, we conducted a *cognitive walkthrough* because we wished to remove as many problems with the planned interface as possible before testing it with users, which was recommended by Lewis and Rieman (1993), Wharton et al. (1994), and others. Applying a *cognitive walkthrough* before conducting a study with users helps to avoid users wasting their time on trivial bugs that could have been caught earlier. Thus, it helps to build the respect of users for the researcher, and results in users giving the design effort serious attention, according to Lewis and Rieman (1993). The results we obtained by conducting the *cognitive walkthrough* were beneficial to us because we were able to adapt the interface design and the respective paper mockups for further evaluation steps. As Lewis and Wharton (1997: 728-29) mentioned, a major challenge of conducting a *cognitive walkthrough* for the evaluator (i.e., the author of this thesis) is to simulate the decisions and thoughts of target users. In our project, this was particularly challenging because the background of the target users (i.e., historians) and the background of the evaluator (i.e., GIScience) are different. Therefore, getting to know the users in the *focus group* was crucial to be able to simulate the thoughts and decisions.

Then, we evaluated the revised interface design in a *think aloud study* with users because we wished to analyze whether target users interact with the planned user interfaces as expected, following the method proposed by Lewis and Rieman (1993). We presented

the tasks and the paper mockups that we adapted based on the findings of the *cognitive walkthrough* to target users. The comments of the participants on decisions and thoughts while solving the tasks were video-taped, which helped us to detect why or why not participants executed anticipated actions with the interfaces. We again found several interaction issues that had to be fixed. In addition, target users suggested additional functionalities to be included in future versions of the interface. This might be due to the use of paper mockups, as users might have had the impression that the interface concept was not finished, and thus that changes were still easily possible, which was already discussed for the *focus group* previously. Therefore, the *think aloud study* was beneficial to us, and we adapted the interface concept accordingly.

Next, we utilized the adapted interface concept in the development stage of the prototype implementations. We designed two interactive web interfaces (i.e., *spatialized networks* and *self-organizing map*) based on Shneiderman's (1996) *visual information-seeking mantra*: "*overview* first, *zoom* and *filter*, then *details-on-demand*", which we considered relevant for our project because it is a basic principle of *information visualization* to design interactive tools which support information-seeking processes in large databases. Applying the *visual information-seeking mantra* presented by Shneiderman (1996) to design interactive interfaces further supports the integration of coupled *distant* and *close reading* functionalities in interactive web interfaces for digital text data (Moretti, 2005, Jockers, 2013), which was a major goal of our project as introduced in *Chapter 1 – Introduction*. The *overview first* of Shneiderman's (1996) *visual information-seeking mantra* relates to the *distant reading* idea, introduced by Moretti (2005), to present overall structures and interconnections to interested information seekers. *Zoom* and *filter* and the subsequent access to *details-on-demand* in Shneiderman's (1996) *visual information-seeking mantra* relate to the *close reading* concept (i.e., to provide information seekers access to the raw data source).

Based on the feedback we gathered by participants of the *focus group* and the *think aloud study* as well as on the findings of the *cognitive walkthrough*, we adapted our interface concept and incorporated Shneiderman's (1996) *visual information-seeking mantra* in the *spatialized network interface* as follows: if a user selects the 19th century network of Switzerland, for example, the spatio-temporal relationships of Swiss toponyms are presented to a user in an interactive network visualization. Therefore, a user gets a first impression (i.e., an *overview*) of the hierarchical structures and overall interconnections of the toponym network of Switzerland in the 19th century (i.e., *distant reading*) based on the HDS articles. Then, a user can *filter* by a specific spatial (i.e., *Switzerland*, *Canton of Zurich*) and temporal scale (i.e., 18th, 19th, 20th centuries), and can do a mouseover or click on a node or edge to see additional information (i.e., *details-on-demand*). For example, a user might be interested in *Zürich*'s strongest relationships in the 19th century network of Switzerland and thus clicks on the network node representing *Zürich*. Then, the three strongest toponym relationships of *Zürich* are displayed on a map and additionally in an information window. In the information window, hyperlinks to articles in the HDS (i.e., articles in which *Zürich* and the toponyms that co-occur most often with *Zürich* in

HDS articles) are displayed and allow users to access the raw data source (i.e., *close reading*).

We incorporated Shneiderman's (1996) *visual information-seeking mantra* in the interactive *thematic landscape* as well. Users access the web page with the interface and see an *overview* of the *thematic landscape*. Themes are displayed in a *self-organizing map* in different colors, and for each theme, the theme's title and the number of articles belonging to that theme are depicted. Therefore, the user receives an impression about the thematic information and structures (e.g., which themes are close to one another) in the *thematic contributions* articles of the HDS (i.e., *distant reading*). A user can *zoom* in to the *detail view* of the *thematic landscape* or can enter query terms in the article title search tool to find themes or articles of interest. If a user clicks on an article or has selected an article that was displayed in the article title search tool, a pop-up window appears (i.e., *details-on-demand*). In the pop-up window, a hyperlink to the selected article in the HDS is provided as well as hyperlinks to ten thematically most related articles, giving users direct access to the raw data source (i.e., *close reading*).

Applying Shneiderman's (1996) *visual information-seeking mantra* to develop interactive user interfaces is a common strategy in geoVA. For example, Roth et al. (2015) developed an exploratory user interface depicting criminal activity data in an interactive map, providing spatial, temporal, and thematic query options, as presented in *Subsection 2.3.3*. Users access the interactive interface and see an *overview* of criminal activity data. Then, spatial, temporal, and thematic *filtering* may be applied, and users may access *detailed* information about potential incidents *on demand*. Another example is illustrated by Luo et al. (2014), who incorporate spatialized displays in contrast to Roth et al. (2015) and analyze the international trade network (see *Subsection 2.3.3*). In Luo et al.'s (2014) interface, a network visualization displays relationships between countries based on the amount of trade, and a map visualizes the spatial distribution of the *gross domestic product* (GDP) by countries. Users may access the interlinked network and map view of the interface and study interesting spatial and social relationships by *filtering* interesting relationships. *Details* about selected relationships (e.g., which sub-groups of core and periphery countries exist in the trade network) are presented to a user *on demand*. Our approach is different from such approaches in geoVA, because we applied Shneiderman's (1996) *visual information-seeking mantra* to spatio-temporal and thematic information automatically retrieved from a semi-structured digital text archive in the humanities (i.e., different context). Luo et al. (2014) also presented a coupled interactive network and map visualization approach, following Shneiderman's (1996) *visual information-seeking mantra*, similarly to our interactive *spatialized network interface*, but applied the approach to non-textual input data (i.e., *quantitative trade network data* and *GDP*). Fabrikant et al. (2015) presented an approach to interactively visualize a *self-organizing map*, similarly to our *thematic landscape*, but used as well non-textual input data (i.e., *census data*). Other applications were presented in *Subsection 2.3.3* (e.g., Tomaszewski, 2008, Robinson et al., 2016) which applied the *visual information-seeking mantra* to textual input data, but did not provide spatialized displays in interactive user interfaces in contrast to our approach.

In *digital and spatial humanities* literature, examples of interactive web tools that follow the *visual information-seeking mantra* (Shneiderman, 1996) and provide *distant* and *close reading* functionalities can be found (Jänicke et al., 2015). However, only a few examples combine spatio-temporal and thematic information, automatically retrieved from unstructured or semi-structured text archives in the humanities and present it in visual and interactive displays for further data exploration (Jänicke et al., 2015). Hinrichs et al.'s (2015) work is a notable exception: they visualize co-occurrences of *place names* and *commodities* (e.g., *sugar*, *coal*), retrieved from an unstructured historical text collection, and present this information in an interactive user interface, including temporal filtering options. However, their solution does not include interactive spatialized displays to explore spatio-temporal or thematic structures or interconnections.

Our approach attempts to respond to the research gap illustrated in *Section 2.5*, as it provides a comprehensive framework to create spatializations based on spatio-temporal and thematic data retrieved from a semi-structured text data archive in the humanities and incorporates these spatializations in interactive user interfaces, following Shneiderman's (1996) *visual information-seeking mantra*, and allows target users to access the digital text archive from a coupled *distant* and *close reading* perspective.

To evaluate the two prototype web interfaces (i.e., the *spatialized network interface* and the *thematic landscape*), we applied a combined *utility* and *usability* evaluation. We introduced in *Chapter 1 – Introduction* our goal to provide interested information seekers in the humanities with exploratory and interactive web interfaces that support sense-making and the generation of new insights about spatio-temporal and thematic information and interconnections in unstructured or semi-structured digital text archives in the humanities. Therefore, we were particularly interested in analyzing the insights that participants gained by interacting with the prototype implementations. For this evaluation step, we considered the work of Nelson et al. (2015) as being relevant because Nelson et al. (2015) analyzed insights gained by target users from interacting with a geoVA interface, applied to text input data. We asked participants to solve tasks by interacting with the web interfaces and to comment on their decisions and thoughts while solving the tasks in a *think aloud study*. We recorded participants' interactions with the web interfaces and audio-taped the *think aloud sessions*. We then analyzed these records and listed insights that participants gained during the *think aloud studies*, as described in *Subsection 5.3.3*.

For the *spatialized network interface*, we summarized insights gained by participants and decided to rank them according to North (2006). We followed North's (2006) approach because we wished to rank insights gained by participants according to different characteristics of insights (i.e., *complexity*, *depth*, *unexpectedness*, and *relevance*) and compare them to one another which is supported by North's (2006) *measuring visualization insights* approach. We revealed that many insights are related to either the structure of the toponym networks in different centuries or hierarchical relationships between toponyms or clusters of toponyms. We expected these findings because network spatializations explicitly highlight hierarchical structures as previously discussed (see answer to *Research Question 2*). Even more, some participants compared the network structures to

the map and realized that some toponyms are connected to one another in the network visualization despite the large geographic distances between them in the map. This was surprising for many study participants, and they mentioned that they might follow up on such insights after the study in order to find out more about the unexpected relationships. Therefore, the *first law of cognitive geography* (Montello et al., 2003), which states that people *believe* that close things (i.e., short network distance) are more similar than distant things, and which applies to the network spatializations, provoked that participants revealed unexpected relationships in the toponym networks if distances in the network do not correspond with distances in the map (e.g., short distance in network, but large distance in the map). We further observed that participants searched for spatio-temporal information in the networks following Shneiderman's (1996) *visual information-seeking mantra*, which is unsurprising because we considered the "*overview* first, *zoom* and *filter*, then *details-on-demand"* principle in the interface development and implementation process, as previously discussed. Therefore, participants searched for interesting spatio-temporal relationships in the network spatialization (i.e., *overview first*), selected interesting toponym relationships (i.e., *filtering*), and read articles in which toponyms co-occur (displayed in the information window of the interface) on the HDS website to find explanations why toponyms are related (i.e., *details-on-demand*). Participants thus used the coupled *distant* and *close reading* functionality we provided in the *spatialized network interface* to gain new insights, and thus Jockers' (2013) suggestion to combine *distant* and *close reading* for information seeking in the humanities is supported by our findings.

For the *thematic landscape*, we were particularly interested in whether participants interpret the *self-organizing map* as expected according to the *first law of cognitive geography* (Fabrikant et al., 2006) and thus whether they interpret themes and articles that are located in close proximity (i.e., the *distance-similarity metaphor*) in the *thematic landscape* as being similar. For this reason, we asked target users to search for articles which are about a combination of two neighboring themes (i.e., *Religion* and *Customs & festivals* in Figure 42) in the *thematic landscape* and expected that they would search in the border region of these two themes in the *thematic landscape*, which is detailed in *Subsection 5.3.3*. We observed participants' behavior while interacting with the *thematic landscape* and realized that participants explored the *thematic landscape* by *zooming* in and out to switch between the *detail* and the *overview map*, and by *panning* to move in different locations in the *thematic landscape*. A typical interaction sequence was that participants identified the border region of the two neighboring themes in the *overview map*, zoomed in to this region in the *detail map*, and then used panning to move along the border region to identify interesting articles. Participants clicked on interesting articles, which made the pop-up window appear, and accessed the original articles on the HDS website by clicking on the hyperlinks provided in the pop-up window. Therefore, participants used the interactive interface by following Shneiderman's (1996) *visual information-seeking mantra*, which we expected as we designed the interface according to this mantra, as previously detailed. This implies, similar to the *spatialized network interface*, that participants gained insights by using both *distant* (e.g., identify themes in the *overview map*) and *close reading* (e.g., access and read original articles on the HDS website) options to

gain insights about thematic information in the HDS. In addition, the results we illustrated in *Subsection 5.3.3* show that participants interpreted distance as a metaphor for dissimilarity (i.e., the *first law of cognitive geography* applies) because all but one article that participants found to be relevant for the two neighboring themes (i.e., *Religion* and *Customs & festivals* in Figure 42) are located in the border region of the two themes in the *thematic landscape*. Furthermore, we revealed that only two participants accessed the search tools on the official e-HDS website, although they were allowed to use the e-HDS tools for 2/3 (i.e., 10 minutes) of the total study time. This supports the usefulness (i.e., *utility*) of our interface to find, access, and interactively explore thematic information in the HDS.

We further tested the *usability* of the two prototype implementations, using the *System Usability Scale* (SUS), which assesses the *global usability* of a system with general measures (Brooke, 1996). We decided to incorporate the SUS for our two prototypes because we were interested to assess how usable our implementations are compared to other systems, as the SUS is a common measure to evaluate *usability* and has already been applied to many applications. According to Bangor et al. (2008), an acceptable system should have at least a score of 70, but good products should score in the higher 70s to upper 80s. The *spatialized network interface* reached a score of 78, and the *thematic landscape* scored 81, which indicates that both interfaces would be ready to be published as full releases, considering the *usability* of the interfaces as one release criteria.

The combined *utility* and *usability* evaluation approach we presented in this thesis is a response to numerous calls in *digital humanities* for the systematic evaluation of interactive web interfaces to provide empirical evidence that these systems are delivering what they promise (e.g., Kirschenbaum, 2004, Gibbs and Owens, 2012). Furthermore, our approach illustrates how target users can be involved early on in an iterative user-centered interface design and evaluation process, which has been listed as the one major requirement for the successful development of visualizations, not only in *digital humanities*, but in general (Jänicke et al., 2015).

By presenting our two prototype implementations, we have illustrated how the existing online version of the HDS might be extended with interactive web interfaces that allow access to the HDS articles from a spatio-temporal and thematic perspective. Our approach particularly highlights the potential benefit of combining *distant* and *close reading* options to interactively explore multidimensional data, following Shneiderman's (1996) *visual information-seeking mantra*, whereas the most recent online version of the HDS only provides *full text* and *article title search* options (i.e., *close reading*).

By answering and discussing *Research Question 3*, we have presented a geoVA approach to visualize spatialized information automatically retrieved from a semi-structured digital text archive in the humanities in interactive and exploratory web interfaces. The combination of evaluation methods with and without users turned out to be very beneficial to remove interface design issues before implementing the prototype versions of the interfaces. We based the interface concept on Shneiderman's (1996) *visual information-seeking mantra* (i.e., "*overview* first, *zoom* and *filter*, then *details-on-demand*")

and have illustrated that target users follow this mantra to gain new insights about spatio-temporal and thematic structures and interconnections in the digital text archive. We further have illustrated that target users have used both the *distant reading* as well as the *close reading* functionalities of our web interfaces in their information-seeking process. This supports Moretti's (2005) argument to provide *distant reading* and Jockers' (2013) argument to couple *distant reading* with *close reading* options to gain new insights into spatial, temporal, and thematic information, structures, and relationships, retrieved from digital text archives.

The discussion of the research questions leading this research illuminates how typical GIScience approaches (i.e., *geographic information retrieval* methods, the *spatialization framework*, and *geovisual analytics* approaches) can be combined and applied to an unstructured or semi-structured digital text archive in the humanities to learn about spatio-temporal and thematic structures and interconnections in the humanities. Text data in the humanities are particularly interesting for GIScience because they contain a great deal of spatial, temporal, and thematic information, and most of this information has not been analyzed with spatio-temporal and thematic approaches to date, as illustrated. In this project, we focused on analyzing and integrating the concepts of *space*, *time*, and *attribute*, which are fundamental to GIScience (see *Subsection 2.3.1*) and geography in general (e.g., Peuquet, 2002, Couclelis, 2005). We studied these concepts in another research context (i.e., humanities) and on data which have not been traditionally studied by GIScience (i.e., text data). In particular, we were able to highlight that Tobler's (1970: 236) *first law of geography*, which states that "everything is related to everything else, but near things are more related than distant things", is transferable to text data in the humanities. Therefore, places that are located close together in *geographic space* co-occur often in text documents (i.e., HDS articles in our project). However, we found several examples that contradict this assumption, which has provoked unexpected insights by target users in the *think aloud study*. To provoke such insights, we transferred the *first law of geography* to the *abstract space* (i.e., relative space) and visualized spatio-temporal and thematic structures and interconnections in network visualizations and *self-organizing maps*. In these *abstract spaces*, the *first law of cognitive geography* applies (Montello et al., 2003), and thus data items visualized close to one another are perceived to be more similar (i.e., *distance-similarity metaphor*). Visualizing both the *geographic* and the *abstract space* in the *spatialized network interface* next to one another, and interlinking them, allowed target users to identify such patterns where distances between spatial data items (i.e., toponyms) in the *geographic* and *abstract space* do not correspond. In addition, incorporating *time* in the spatio-temporal networks provoked insights by target users regarding the change of spatio-temporal relationships and hierarchical structures over time. Furthermore, the *thematic landscape* illustrates how the *first law of cognitive geography* (Montello et al., 2003) can be used to present thematic data (i.e., *attributes*) in *abstract space*, applying *spatial metaphors* and thus allowing interested information seekers to access thematic structures and interconnections in large digital text archives in the humanities. Therefore, we proposed answers to what Couclelis (2005: 35-36) defined as major research challenges in GIScience: the integration of *space* and *time* (and *attributes*) in

GIScience, and the representation of *relative* and *non-metric spaces* (and *times*) as a supplement to *absolute* spaces.

To summarize, applying typical GIScience approaches to humanities text data thus does not only support answering (new) research questions in the humanities, but also opens up new research possibilities to study GIScience approaches as well as theories and core concepts of geography (i.e., *space*, *time*, *attributes*) in a research context outside of GIScience and geography.

In the next section, we consider limitations we revealed by applying the GIScience approach demonstrated in this thesis.

7.2 Limitations

In the previous section, we have illustrated and discussed the answers to the research questions of our project. In this section, we focus on limitations we revealed. We first start by limitations that are due to the data sources we employed in this project. Then, we will focus on limitations of the methods we applied.

For this project, we decided to apply the GIR algorithm developed by Derungs and Purves (2014) to retrieve spatial information from the HDS. This algorithm fits the purpose of our project very well because it is optimized for retrieving toponyms from unstructured German text documents, as detailed in *Subsection 4.1.1*. The algorithm employs *SwissNames* as a gazetteer to retrieve toponyms from text documents. Selecting *SwissNames* has an influence on the retrieval results. First, *SwissNames* only covers the region of Switzerland as it is today, and only contains toponyms which are mentioned on a topographic map of Switzerland. Therefore, all places that are located outside of Switzerland cannot be retrieved. This has an influence on the visualization of the *spatialized networks* of toponyms. Consequently, target users can only gain insights about spatio-temporal relationships of Swiss toponyms, whereas all relationships from Swiss toponyms to the region outside of Switzerland are not depicted in the networks. This is relevant because the toponym relationship networks ignore the fact that Switzerland has strong relationships to other countries and regions outside of Switzerland. For example, large parts of the German-speaking region of Switzerland were part of the *Diocese Constance* (= *Diozöse Konstanz*) until 1815, which is documented well in the HDS (Bischof and Maurer, 2016). However, all toponym relationships from Swiss cities or villages to *Constance* are not depicted in the network spatializations in this thesis, as *Constance* is a German city.

Second, spatial divisions on the district (= *Bezirk*) or canton (= *Kanton*) level, or other levels higher than cities, villages, and municipalities, are not considered in *SwissNames*. As described in *Subsection 4.1.1*, we only considered the cantons for our study in the disambiguation process of the spatial information retrieval algorithm. Consequently, the term *Bezirk Winterthur* (= district of Winterthur), for example, would be recognized as the city instead of the district of *Winterthur* by the spatial information retrieval algorithm.

This is relevant for the visualization of the *spatialized networks*, because the toponym *Winterthur* is depicted more central than it would be if we would have separated the toponym for the *Bezirk Winterthur* from the toponym for the city of *Winterthur*. Therefore, all cities or villages that have the same name as a *Bezirk* are visualized more central in the network spatializations compared to cities or villages of which no *Bezirk* with the same name as the city or village exists.

Furthermore, we illustrated in *Section 3.3* that many people are involved in the writing and publishing process of the HDS articles (e.g., *authors*, *scientific advisors*, *editorial office*), and the HDS responsible people took decisions about which *themes*, *geographical entities*, *people*, and *families* are covered in the HDS. This is relevant for our thesis, as these people and decisions regarding the content of the HDS influence the spatio-temporal and thematic content in the HDS, and thus the contents in the web interfaces we have implemented. Therefore, the results we presented in this thesis are limited to the view of the HDS on Swiss history at the end of the 20th and the beginning of the 21st century. Other findings might result if another data set about Swiss history would have been studied. However, we did not evaluate that in the context of this thesis.

Having illustrated limitations that are due to the selected data sources for this project, we now turn to limitations regarding the methods we applied. As shown in *Chapter 6 – Evaluation*, while we did evaluate the chosen parameters for the *spatialized networks*, we did not evaluate parameter choices for the GIR algorithms. We applied existing GIR algorithms in this project (for example, the method developed by Derungs and Purves (2014) to automatically retrieve toponyms from the HDS articles). For the spatial, temporal, and thematic GIR algorithms, we chose standard parameters, but did not evaluate potential effects on our results when changing these standard parameters. We were able to reach an acceptable level of precision for the spatio-temporal GIR results (see *Section 6.1* and *Section 7.1*). We also illustrated that the automatic clustering of HDS articles corresponds well with human annotated judgments of similarity (see *Section 6.4* and *Section 7.1*). However, systematic evaluations to test the GIR parameters, similar to the assessment of parameters for the computation of the *spatialized networks* (see *Section 6.2*), would be necessary to test potential optimizations of the GIR results.

A further limitation is related to the automatic retrieval of spatial information. In *Subsection 2.1.1* we have reported about the existence of *metonyms*. Metonymically used toponyms possess a meaning different from the literal, geographic sense (Leveling and Hartrumpf, 2008). Therefore, metonyms do not refer to a location, but to another named entity. One example is the named entity *Bern*: *Bern* either refers to the capital of Switzerland (i.e., *Bern* is used as a toponym), or to the Swiss government (i.e., *Bern* is metonymically used), depending on the context. Leveling and Veiel (2007: 902) found that 17% of the location names in a German newspaper corpus are used metonymically. This is comparable to our project: the author of this thesis randomly selected ten HDS articles in a small case study and found that 19% of the toponyms are metonymically used according to the definition of metonyms provided in Leveling and Hartrumpf (2008: 291-92). However, we have not considered methods that allow to differentiate

metonyms from toponyms in the HDS articles. This might have an influence on the precision of the spatial information retrieval results in the spatialized networks (see *Section 6.1*). We found an overall precision of 83% for the retrieval of spatio-temporal information from HDS articles. By applying approaches to efficiently separate metonyms from toponyms, we would expect the precision to increase. This, however, would need to be systematically evaluated. Suggested approaches to automatically detect metonyms in text documents are presented in *Section 8.3*.

Another limitation we highlight is related to certain core assumptions made in this project: we have assumed that a relationship between two toponyms exists if they co-occur often in the same HDS articles. This assumption is based on prior work by Hecht and Raubal (2008) and Salvini and Fabrikant (2016), and is based on a *bag of words* approach. *Bag of words* in this context means that an article text is considered as one *bag* of its *words*, and the syntax of the text documents (e.g., word order, sentence structure) is disregarded. Gregory and Hardie (2011) and Murrieta-Flores et al. (2015) illustrate how the syntax could be considered to relate toponyms and semantic concepts (e.g., the term *war*) in digital text documents to one another by analyzing the textual proximity of toponyms and concepts: the more often a toponym and a concept are mentioned in close (textual) proximity (e.g., in the same sentence), the stronger the relationship between the respective toponym and concept. Similarly, the relationship between two toponyms could be assessed by using the distance in words between them (e.g., the lower the textual distance between two toponyms, the stronger the relationship). However, in our project we did not cover such syntax-based approaches, and thus cannot illustrate whether incorporating the syntax would change the precision of spatio-temporal information in the spatialized networks (see *Section 6.1*). Systematic evaluations would be necessary to be done to directly compare the results of a *bag of words* with a syntax-based information retrieval approach. As a further assumption, we have defined that HDS articles are only relevant for toponym relationships if at least 50% of the temporal references in an article refer to a specific century (see *Subsection 4.2.1*). We have shown in the evaluation section of this thesis (i.e., *Section 6.2*) that this can have a significant influence on the contribution of the *geographical entities*, *families*, and *thematic contributions* article categories to the total strength of weighted toponym relationships in the networks, because many articles in these categories are excluded from the computation of the toponym networks by applying the *50% criterion*. Therefore, the contribution of the aforementioned article categories is low compared to if the *50% criterion* would not be applied. This is because articles in these categories typically cover extended time periods, and thus only a few articles fulfill the *50% criterion*, compared to articles of the *biographies* category, which are typically more easily attributable to single centuries (see *Section 6.2*). We did evaluate the influence of adapting the *50% criterion* on the network visualizations, and report that adapting the criterion has no major influence on the general network structure (see *Section 6.2*). Nevertheless, we can expect that some articles might be excluded, even though they might be relevant for specific centuries. For example, the long HDS article about *agriculture*, which covers an extended time period, is not contained in any toponym network. This article does contain 17 temporal references about the 19th century

(i.e., 28.3% of all temporal references in this article), which indicates that this article could indeed be relevant for the 19th century. To investigate whether such kinds of articles are relevant for target users to understand and explore the spatio-temporal toponym relationships when interacting with the *spatialized network interface*, we would need to conduct further evaluations to determine whether parameters or the method to compute toponym relationships needed to be adapted.

Apart from the assumptions related to GIR, the visualization of the spatialized information required for additional decisions. We provided arguments for the network visualization and SOM techniques in *Subsection 2.2.3*. We particularly highlighted the strength of network visualizations to visually emphasize relationships between data items and to graphically depict hierarchical structures and the centrality of nodes within the networks. Our choice for SOMs to show thematic information and structures, was mainly based on the strength of SOMs to highlight semantic relatedness by placing similar data items close to one another in a map-like visualization. This facilitates the identification of clusters of semantically related objects. We did, however, not compare the influence of the selected visualization techniques on the process of gaining insights by target users of our project. For example, we did not assess whether target users gain different insights by interacting with the interactive web interfaces compared to reading article texts only. We also did not assess, whether the selected parameters for the creation of visualizations (e.g., *graph embedder* layout / *minimum spanning tree* for the spatialized networks) have an influence on the process of gaining insights. Furthermore, we did not evaluate whether and how the use of other visualization types (than networks and SOMs) would influence the target users and their information-seeking process. For example, tree map visualizations also emphasize hierarchical structures in large data sets, and thus could be tested as an alternative to spatialized networks. Tree maps, however, are not as powerful in highlighting relationships between single data items compared to networks (Shneiderman, 1996). It would be interesting to analyze how such characteristics of visualization types influence target users in gaining insights. However, we did not test this in our research project.

We identified a further limitation, which is related to the prototype implementation of the *spatialized network interface*. We incorporated the spatio-temporal networks of the 18th, 19th, and 20th centuries at the country and *Canton of Zurich* level, including the 203 toponyms that occur most often in the HDS, which was discussed in *Subsections 4.2.1, 4.3.2,* and *5.1.1*, respectively. These decisions are relevant for the reported results, because information seekers interested in Swiss history are limited to gaining insights to these pre-selected spatial and temporal information that are displayed in the *spatialized networks*. For example, one consequence was that participants spent significantly more time looking at the networks of *Switzerland* compared to the networks of the *Canton of Zurich*, in the second *think aloud study*, as reported in *Subsection 5.3.3*, which was primarily due to the limited information and content (i.e., small networks with only a few nodes and edges) in the *Canton of Zurich* networks, according to participants. Turning now to the temporal information, we chose to present spatio-temporal relationships in a static view on the century level. Therefore, users of

the *spatialized network interface* did not have the possibility to aggregate the spatio-temporal information on another temporal scale such as *decades* or *years*. As we reported in *Subsection 4.2.1*, selecting a higher temporal resolution than *centuries* would result in incomplete networks as there is not much information available in the HDS about most of the 203 Swiss toponyms for many *decades* and individual *years*. This is then a limitation due to the limited size of the selected data source for this project (i.e., HDS). A further limitation is that users cannot select a *time range* (e.g., 1939-1945) useful and meaningful for them to study Swiss history, and thus cannot study spatio-temporal relationships for a particular historical event (e.g., *Second World War*). Furthermore, in the *spatialized network interface*, users can only view spatio-temporal relationships in a selected century. Additionally, changes over time (e.g., strength of relationship that changes over time) are not explicitly visualized (e.g., using a visual variable) or highlighted in the prototype implementation. Potential future work to resolve these limitations are presented in *Section 8.3*.

The last limitation we report is related to the results of the second *think aloud study* we conducted to test the *utility* and the *usability* of our prototype implementations. As we reported in *Subsection 5.3.3*, due to the low number of participants (i.e., five participants) in the *think aloud sessions*, we were not able to conduct statistical tests of the SUS scores and of the questions we asked participants whether they were satisfied with the results they obtained, how confident they were that they reached the goal of the task, and how relevant they thought their insights were regarding the history of Switzerland. This was relevant for our project because we only could compare these scores descriptively (i.e., the average score) to one another instead of applying statistical tests to compare them. To publish a full release of the interfaces, which have to be optimized for *usability*, we would need to test *usability* with a larger sample of participants to conduct quantitative studies (i.e., statistical testing). Nielsen (2012) suggests to aim for about 20 users for such *quantitative usability testing*.

In this chapter, we discussed the results of our research project and related our findings to the state of the art in research fields relevant to this project. In the second section of this chapter, we illustrated limitations of our approach, which are related to the applied methods as well as the data sources we employed. We will illustrate the achievements and contributions of our project based on the findings of this chapter and present ideas for future work in the next chapter.

8 Conclusions and Outlook

To conclude this research project, we reflect on the results and the discussion of the results of this thesis, and point out the main achievements in *Section 8.1*. Then, based on these achievements, in *Section 8.2* we present our contribution to the different research fields that are relevant to this research project. Based on the limitations we revealed in *Section 7.2*, we illustrate potential future work in *Section 8.3*.

We start this chapter by highlighting the main achievements of this thesis.

8.1 Achievements

This research project was motivated by the wealth of information that is online in large unstructured and semi-structured text archives, and the limited mechanisms to assess and depict knowledge in these archives in order to allow information seekers to gain new insights about the text documents. Unstructured and semi-structured text archives in the humanities were found to be particularly relevant in the context of this thesis, since they contain a great deal of spatial, temporal, and thematic information, largely untapped for spatio-temporal and thematic analyses in geography to date. Making such information available to target users in cognitively supportive and perceptually salient user interfaces to gain new insights and find new research hypotheses about space, time, and theme in the humanities is a major research challenge and was addressed in this thesis adopting a GIScience perspective.

To address this challenge, we introduced three main goals of this thesis in *Chapter 1 – Introduction*: we first wished to retrieve spatio-temporal and thematic information automatically from a large unstructured or semi-structured digital text archive in the humanities such that hidden structures and relationships can be uncovered. We chose the *Historical Dictionary of Switzerland* (HDS) as a case study, as one typical, digitally available semi-structured text archive in the humanities. Second, we aimed to visualize uncovered spatio-temporal and thematic structures and relationships in two-dimensional spatialized displays that allow further data exploration. Third, we wished to make these spatialized displays available in interactive and exploratory web interfaces that support sense-making and the generation of new insights to interested information seekers in the humanities. Reaching these goals in a comprehensive approach has been missing in

literature to date. Thus, we identified it in *Section 2.5* as a research gap for this thesis. To bridge the identified research gap, we suggested a combined *geographic information retrieval*, *spatialization*, and *geovisual analytics* approach.

We first demonstrated how existing *geographic information retrieval* methods can be applied to automatically retrieve spatial, temporal, and thematic information from the HDS. We illustrated that the HDS contains a wealth of spatio-temporal and thematic information. The retrieved information was transformed and reorganized such that spatialized displays could be generated in the next step. We depicted relationships between Swiss toponyms in different time periods in *network spatializations* and illustrated that networks support information seekers in gaining insights into spatio-temporal patterns and hierarchical structures of the toponym network. In addition, we chose *self-organizing maps* to depict the 3,067 HDS *thematic contributions* articles and demonstrated that *self-organizing maps* support information seekers in identifying themes and articles that are covered by the HDS. We then presented a user-centered evaluation and interface design approach to develop interactive web interfaces that incorporate the *network spatializations* and the *self-organizing map*. We highlighted the benefit of involving target users (i.e., historians) early on in the interface design process to assess the needs and requirements of our target group and to remove potential interface design issues before implementing interactive web interfaces. In the next step, we developed a *spatialized network interface* that allows information seekers to study spatio-temporal relationships in three different time periods (i.e., 18th, 19th, and 20th centuries) and at two different spatial scales (i.e., *Switzerland* and *Canton of Zurich*). In addition, we represented the *thematic contributions* articles of the HDS in an interactive *thematic landscape* (i.e., *self-organizing map*) that allows information seekers to search for interesting themes and articles covered by the HDS text archive. Finally, we evaluated our results in a combined *utility* and *usability* study and demonstrated that target users were able to gain new insights into spatio-temporal and thematic structures and interconnections in the HDS by interacting with the two web interfaces. Therefore, we were able to reach the goals we formulated for this thesis and were able to learn and gain insights into space, time, and theme in the humanities by applying a typical GIScience approach.

In the next section, we illustrate to which research fields we contribute with our approach and the results we achieved in this research project.

8.2 Contributions

In this section, we put the results and discussion of the results of this thesis into the scientific context and highlight our contributions from two perspectives: a GIScience/geography perspective, and a *digital humanities* perspective. We start with the former.

We contribute to *geovisual analytics* (geoVA) as we suggest a methodological framework to gain new insights about spatio-temporal and thematic structures and interconnections in unstructured and semi-structured digital text archives in the humanities. As we

illustrated in *Section 7.1*, our approach is different from previous approaches in geoVA, as it suggests a systematic framework incorporating all steps from raw text data processing, the automatic retrieval of spatial, temporal, and thematic information from unstructured and semi-structured text documents, the visualization of this information in spatialized displays, to the incorporation of these spatializations in interactive web interfaces, involving target users in the interface design and evaluation process. Therefore, our approach might serve the geoVA community for future work on gaining insights from unstructured or semi-structured text archives (in the humanities) about space, time, and theme.

We further contribute to *geographic information search*. As we introduced in *Chapter 1 – Introduction*, Ballatore et al. (2015: 6) pointed out at a specialist meeting about *Spatial Search* that "spatial, temporal, and thematic information are always interacting in search(…)". Therefore, "(…) search should be able to integrate and combine these dimensions" (Ballatore et al., 2015: 12). This was emphasized by Grossner (2014: 26), who stated that "spatial searches often require or benefit from contextualizing temporal and thematic parameters". The two interfaces we implemented in the course of this project both provide multidimensional (i.e., spatial, temporal, and thematic) search capabilities: the *spatialized network interface* allows spatial and temporal search and displays thematic information (i.e., articles in which toponyms co-occur) on demand in information windows. The *thematic landscape* provides search functionalities for thematic data, visualized in a spatialized view (i.e., *self-organizing map*). Furthermore, Ballatore et al. (2015) highlight the importance of evaluating exploratory search in the context of unstructured data in *geographic information search* and state that "…geography can benefit from new search approaches to explore data and formulate new research questions" (Ballatore et al., 2015: 15). We demonstrated an iterative user-centered design and evaluation approach for a semi-structured text archive. Furthermore, participants in our user studies revealed facts interesting and unexpected to them by interacting with the interfaces and even declared that they might do more research on insights they gained while interacting with the interface (e.g., on Insight 1 in Table 18). This supports the *utility* of the *geographic information search* functionalities we implemented in the two web interfaces for information seekers and might inspire future research in *geographic information search* to incorporate combined spatio-temporal and thematic search functionalities.

We illustrated in *Section 7.1* how our research project might support the understanding of fundamental concepts of geography and GIScience (i.e., space, time, and attributes), analyzed in research fields outside of geography (i.e., text documents in the humanities). We further promote our research approach to other fields in geography, such as *urban geography*. Salvini (2012) depicted the *global city* network in *spatialized networks* by analyzing the hyperlink structure on *Wikipedia*. Our approach differs from Salvini (2012) because we analyzed the toponym network on a national and regional level, and additionally incorporated the temporal dimension, as discussed in *Section 7.1*. We showed that incorporating the temporal dimension allows information seekers to uncover facts about changing (historical) spatial dependencies and interconnections. Such insights

might inspire urban geographers to generate new research questions about the dynamic relationships of toponyms over time.

Having illustrated our main contributions to GIScience/geography, we now turn to the *digital humanities* perspective. Our approach contributes to the *digital humanities* community as this community seeks new ways to expand the humanities with computing technologies, and fosters interdisciplinary research collaborations between humanities experts and experts in modern digital technologies (Kaplan, 2015). The need for geographic approaches in *digital humanities* is highlighted by several research fields at the nexus between geography and the humanities, such as *GeoHumanities*, *spatial humanities*, and *historical GIS* (see *Subsection 2.4.2*), which all seek to connect typical theories, concepts (e.g., *space*, *time*), and approaches in geography with the humanities.

With this research, we contribute a combined spatio-temporal and thematic information retrieval approach to this community. This approach allows users to automatically identify and extract spatio-temporal and thematic information from large unstructured or semi-structured digital text archives in the humanities such that spatio-temporal and thematic structures and relationships can be analyzed. We further test the systematic and theory-driven application of the *spatialization framework* (Fabrikant and Skupin, 2005) to spatio-temporal and thematic information retrieved from unstructured or semi-structured digital text archives to *digital humanities*. The visualization of the spatialized information retrieved from large text data archives in lower dimensional representations (i.e., *network visualizations* and *self-organizing maps*), informed by the empirically tested *spatialization theory*, seems to be useful because participants in our user studies were able to gain new insights about spatio-temporal and thematic structures and interconnections in the text archive by interacting with prototype web interfaces of the spatialized displays (i.e., *spatialized network interface*, *thematic landscape*). We further presented an interface design approach that is based on Shneiderman's (1996) *visual information-seeking mantra* and provides both *distant reading* and *close reading* functionalities (Moretti, 2005, Jockers, 2013). We were able to show that participants made use of both the *distant* and *close reading* functionalities in the interactive spatialized displays to gain insights, and indeed followed the "*overview first, zoom and filter, then details-on-demand*" strategy (Shneiderman, 1996) to interact with the web interface. Finally, we contribute the systematic user-centered evaluation and design methodology we applied in our project to *digital humanities*. The approach we presented in this thesis, which involved several steps with and without target users, is useful for learning about needs and requirements of target users in the humanities and helps to remove potential interface design issues, as demonstrated in this thesis.

We have presented an approach at the nexus of GIScience and *digital humanities*. However, our approach might be employed outside academia as well. For example, many economic branches that deal with large text data might benefit from applying an approach similar to that presented in this thesis. For example, telecommunication companies or airlines might analyze digital customer feedback with our approach to structure the feedback by space, time, or theme, and thus to filter information which is relevant to them. An example for such an application is presented by Janetzko et al.

(2014), who spatio-temporally clustered and visualized customer feedback in a *self-organizing map*.

Now that we have illustrated the achievements and contributions of our research project, we will turn to the outlook and present potential future research directions.

8.3 Outlook

We illustrated an approach at the nexus of GIScience and the humanities. At this nexus, we review potential future research directions, and we further provide ideas for future research which might overcome limitations we uncovered in *Section 7.2*.

- **Refine the spatio-temporal relationships computation**. The algorithm to compute the spatio-temporal relationships for the *spatialized network interface* is based on toponym co-occurrences and a temporal weight factor, regardless of the position of the spatial and temporal information in the text (i.e., *bag of words* approach). In future work, the position of the spatio-temporal information in the text might be used to generate an additional weight for the toponym relationship calculations. For example, if toponyms and temporal references co-occur in the same paragraph of an article, the weight of the spatio-temporal relationship of these toponyms might be increased.
  In addition, the relevance of articles considered for the computation of *spatialized networks* might be further discussed and evaluated with target users in the humanities in future studies. Results of such evaluations might help to optimize parameters (e.g., the *50% criterion* for temporal references), and thus to increase the number of spatio-temporal relationships that are deemed relevant by target users, and thus need to be presented to them in the *spatialized network interface*.

- **Evaluate and test parameter choices and assumptions: GIR**. We assessed parameters for the *spatialized networks* and evaluated the influence of parameter choices on the spatialized displays. However, parameters employed for the applied GIR algorithms have not been systematically assessed and evaluated in this thesis, and is recommended for future work. Results of such evaluations might support the optimization of the GIR results, and thus might help to improve the quality (e.g., *precision* of spatio-temporal retrieval results) of the spatialized displays. Another step towards the improvement of the GIR results, would be the development of a spatial information retrieval method that allows the separation of the metonymical use of place names in text documents from place names that refer to a location (see *Section 7.2*). In literature, several approaches are suggested to achieve this goal. For example, Leveling and Hartrumpf (2008) suggest a language-independent machine learning solution that considers the textual context in which place names are used as one factor to automatically identify metonyms. They illustrate that the precision of the spatial information retrieval results can be substantially improved, using the approach they suggest. A similar approach could be applied to the HDS or other text repositories in the humanities that could

evaluate whether applying such approaches influences the precision of the spatial information retrieval results.

- **Evaluate different visualization techniques**. In future work, different visualization types (e.g., network visualizations and tree maps) could be incorporated in the interactive user interfaces. If various visualization types are available to target users to study the same information, the assessment of whether and how these visualization types influence the target users' insights can be achieved. The assessment of whether the process of gaining insights differs for the selected visualization techniques would also be possible. Based on the results of such user studies, suggestions for the use of particular visualizations techniques and parameters for specific tasks could be elaborated, and provided to the digital humanities and information science community.

- **Apply to other language versions/data sources**. We suggest the application of the presented methods to other language versions of the HDS as well as to other text archives in the humanities. First, the application to other language versions of the HDS would allow the detection of potential differences in the results due to the methods applied, as the content in all language versions is equal. Therefore, if comparisons of different language versions would be applied, it would be possible to draw conclusions about the stability of the methods we employed and the potential applicability of our approach to text documents in other languages. Applying the methods presented in this project to other language versions poses challenges for the GIR algorithms we employed. For example, to retrieve spatial information from HDS articles in another language than German (i.e., French or Italian), gazetteers would need to be adapted (e.g., using a gazetteer with Swiss toponyms in French or Italian). In addition, the text preprocessing algorithms (e.g., exclusion of abbreviations) would need to be substantially changed for a different language version. Second, it would be useful to apply the methods presented in this thesis to another data source describing the history of Switzerland. This would allow us to assess whether spatio-temporal structures and interconnections revealed by analyzing the HDS could also be found in other data sources. This would enable us to state whether found patterns are due to the view of Swiss history by the people involved in the writing and publishing process of the HDS articles only, or if the findings are valid for other views on Swiss history as well. Potential data sources to perform such a comparison could be ancient encyclopedias about Swiss history such as the AHESL or the HBLS (see *Section 3.1*). However, applying the approach to historical text collections is challenging, because the methods we applied in this project are optimized for modern language. For example, the gazetteer we applied (i.e., *SwissNames*) to retrieve spatial information from the HDS only contains toponyms to describe places that are used today, and thus ancient spellings of toponyms would be missed in historical text collections. The use of historical gazetteers may be one solution to address this challenge, but this is not further discussed here, and interested readers are referred to Southall et al. (2011). Third, the application of the presented approach might be tested with

other data sources in the humanities to test the transferability and generalizability of our approach. Furthermore, applying the approach to non-humanities text archives (e.g., *economy*) is another possible extension for future work. It would be particularly interesting to analyze how well the approach performs on data sets which contain a small amount of spatio-temporal and thematic information in contrast to the data set we investigated in this project and to analyze whether the results in other contexts would still be reasonable and useful for target users.

- **Extend spatio-temporal information access functionality**. For this project, we analyzed the spatio-temporal relationships of toponyms in *Switzerland* and in the *Canton of Zurich* in the 18th, 19th, and 20th centuries. In future projects, both the considered spatial as well as temporal scale might be extended. For example, further cantons of Switzerland might be incorporated in the *spatialized network interface* to compare spatio-temporal structures and interconnections in different cantons to one another. Furthermore, spatio-temporal relationships of Swiss toponyms to toponyms outside of Switzerland could be analyzed by employing another gazetteer (e.g., *GeoNames*, see *Subsection 2.1.1*) for the spatial information retrieval other than *SwissNames* (see *Section 7.2*). This would allow information seekers to inspect which countries and regions around Switzerland are most related to Switzerland and how such spatial relationships changed over time. Furthermore, the *spatialized network interface* functionalities could be extended such that time ranges interesting for target users could be freely selected (e.g., 1939-1945). This could be achieved by, for example, providing a dynamic time slider in the interactive user interface. The user interface could be implemented such that it dynamically assesses all relevant spatio-temporal relationships (i.e., for the selected time range) from a spatio-temporal database, using a database query language (e.g., PHP/SQL or d3.js/SQL implementation). Based on the output of the query, a network displaying the spatio-temporal relationships for a selected time period could then be dynamically rendered in the user interface.
  In addition, changes over time could be highlighted in the *spatialized network interface*. To visualize the increasing or decreasing strength of a relationship between two toponyms over time, the visual variable *color value* could be used. For example, if the relationship between two toponyms is very weak in the 18th century, but very strong in the 19th century, this could be visualized by using a dark color value for the edge that connects the two respective toponyms in the 19th century. In contrast, relationships between toponyms that get weaker over time could be visualized using a light color value. Alternatives to depict changes over time would be to visualize the spatio-temporal relationships in animated displays (i.e., node and edge size dynamically change over time) or in flow diagrams. Flow diagrams could for example be used in order to show the changing constitution of toponym communities (i.e., clusters of toponyms that co-occur often in the same HDS articles) over time. Such visualization types could be displayed in a pop-up window by demand instead of incorporating them in the existing user interface. This could decrease the overload of information presented to target users.

- **Thematic landscape at primary, secondary, and high schools**. Participants of the second *think aloud study* suggested promoting the *thematic landscape* interface to schools because they expected that pupils might be additionally motivated to gain new insights into the history of Switzerland by interacting with the interactive web interface as a supplemental tool to more traditional learning resources (e.g., books). This idea was supported by a professor of media pedagogy at a Swiss *teacher training college* (= *Pädagogische Hochschule*) in an informal interview after completing the empirical evaluations of the interfaces. According to this expert, *collaboratively working* should be supported by such an interface, and *didactic instruments* for teachers with guidelines on how the interface can be used in classroom exercises should be provided. This might be a future research niche that could be occupied by an interdisciplinary research collaboration of media pedagogy experts, history teaching experts, and GIScientists.

We began this thesis by reporting about the wealth of information that is stored in large online text archives. We illustrated our contribution to the discussion about how new insights can be gained from large text collections. We particularly highlighted how spatio-temporal and thematic information can be automatically retrieved from digital text archives in the humanities; how this information can be transformed, reorganized, and interactively presented to target users; and how this might help to explore and learn about spatio-temporal and thematic structures and relationships buried in the text archives. We thus provided a typical geographic answer for how information access to and the exploration of large online text archives can be facilitated. With this thesis, we aim at inspiring future research projects situated at the nexus of geography and the humanities dealing with the visualization and exploration of spatio-temporal and thematic information in large digital text archives.

References

Adams, B & Gahegan, M 2016 Exploratory Chronotopic Data Analysis. in **Miller, J A, O'Sullivan, D & Wiegand, N** eds *Geographic Information Science: 9th International Conference, GIScience 2016, Montreal, QC, Canada, September 27-30, 2016, Proceedings*. Springer International Publishing, Cham, Switzerland 243-58.

Ahn, D, van Rantwijk, J & de Rijke, M 2007 A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. *Proceedings of the HLT-NAACL 2007*. Rochester, NY, USA.

Albert, R & Barabási, A-L 2002 Statistical mechanics of complex networks. *Reviews of modern physics* 74 47-97.

Alex, B, Byrne, K, Grover, C & Tobin, R 2015 Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing* 9 (1) 15-35.

Alonso, O, Gertz, M & Baeza-Yates, R 2007 On the value of temporal information in information retrieval. *SIGIR Forum* 41 (2) 35-41.

Alonso, O, Strötgen, J, Baeza-Yates, R & Gertz, M 2011 Temporal Information Retrieval: Challenges and Opportunities. *Proceedings of the 1st international temporal web analytics workshop (TWAW 2011)*. Hyderabad, India 1-8.

Amitay, E, Har'El, N, Sivan, R & Soffer, A 2004 Web-a-Where: Geotagging Web Content. in **Sanderson, M, Järvelin, K, Allan, J & Bruza, P** eds *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, Sheffield, UK 273-80.

Andrienko, G, Andrienko, N, Bremm, S, Schreck, T, Von Landesberger, T, Bak, P & Keim, D 2010a Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. *Computer Graphics Forum* 29 (3) 913–22.

Andrienko, G, Andrienko, N, Demsar, U, Dransch, D, Dykes, J, Fabrikant, S I, Jern, M, Kraak, M-J, Schumann, H & Tominski, C 2010b Space, time and visual analytics. *International Journal of Geographical Information Science* 24 (10) 1577-600.

Andrienko, G, Andrienko, N, Jankowski, P, Keim, D, Kraak, M J, MacEachren, A & Wrobel, S 2007 Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science* 21 (8) 839-57.

Anselin, L 1989 What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis. *Technical Report 89-4*.

Anselin, L 1995 Local Indicators of Spatial Association—LISA. *Geographical Analysis* 27 (2) 93-115.

Anselin, L 2003 Spatial Econometrics. *A Companion to Theoretical Econometrics*. Blackwell Publishing Ltd, Malden, MA, USA 310-30.

Anselin, L, Syabri, I & Kho, Y 2010 GeoDa: An Introduction to Spatial Data Analysis. in **Fischer, M M & Getis, A** eds *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg 73-89.

Archambault, D, Purchase, H & Pinaud, B 2011 Animation, Small Multiples, and the Effect of Mental Map Preservation in Dynamic Graphs. *IEEE Transactions on Visualization and Computer Graphics* 17 (4) 539-52.

Archambault, D & Purchase, H C 2012 The mental map and memorability in dynamic graphs. *Proceedings of the 5th PacificVis Symposium (Pacific Visualization)*. Songdo, South Korea 89-96.

Archambault, D & Purchase, H C 2013 Mental Map Preservation Helps User Orientation in Dynamic Graphs. in **Didimo, W & Patrignani, M** eds *Graph Drawing: 20th International Symposium, GD 2012, Redmond, WA, USA, September 19-21, 2012, Revised Selected Papers*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Aurenhammer, F 1991 Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys (CSUR)* 23 (3) 345-405.

Ayers, E L 2010 Turning toward Place, Space, and Time. in **Bodenhammer, D J, Corrigan, J & Harris, T M** eds *The Spatial Humanities - GIS and the Future of Humanities Scholarship*. Indiana University Press, Bloomington, IN, USA 1-13.

Bach, B, Pietriga, E & Fekete, J-D 2014 GraphDiaries: Animated Transitions and Temporal Navigation for Dynamic Networks. *IEEE Transactions on Visualization and Computer Graphics* 20 (5) 740-54.

Ballatore, A, Hegarty, M, Kuhn, W & Parsons, E 2015 Spatial Search, Final Report. https://escholarship.org/uc/item/33t8h2nw (accessed April 2016).

Bandelier, A, Froidevaux, P, Prongué, J-P & Rebetez, J-C 2009 Basel (Fürstbistum). *Historical Dictionary of Switzerland (HDS)*. http://www.hls-dhs-dss.ch/textes/d/D8558.php (version date 2009/09/10, accessed August 2016).

Banerjee, M 1999 Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics* 27 (1) 3-23.

Bangor, A, Kortum, P T & Miller, J T 2008 An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction* 24 (6) 574-94.

Barbieri, N, Manco, G, Ritacco, E, Carnuccio, M & Bevacqua, A 2013 Probabilistic topic models for sequence data. *Machine Learning* 93 (1) 5-29.

Barker, E, Pelling, C, Bouzarovski, S & Isaksen, L 2010 Mapping the World of an Ancient Greek Historian: The HESTIA Project. *Proceedings of the Digital Humanities 2010 conference*. London, UK.

Baumann, W & Moser, P 2012 Agrarpolitik. *Historical Dictionary of Switzerland (HDS)*. http://www.hls-dhs-dss.ch/textes/d/D13789.php (version date 2012/08/16, accessed September 2016).

Becker, R A, Eick, S G & Wilks, A R 1995 Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics* 1 (1) 16-28.

Behrens, N, Motschi, A & Schultheiss, M 2015 Zürich (Gemeinde). *Historical Dictionary of Switzerland (HDS)*. http://www.hls-dhs-dss.ch/textes/d/D171.php (version date 2015/01/25, accessed August 2016).

Bender-deMoll, S & McFarland, D A 2006 The Art and Science of Dynamic Network Visualization. *Journal of Social Structure (JoSS)* 7 (2).

Bertin, J 1967 *Sémiologie Graphique: Les Diagrammes - les Réseaux - les Cartes*. Mouton, Paris.

Berzak, Y, Richter, M, Ehrler, C & Shore, T 2011 Information Retrieval and Visualization for the Historical Domain. in **Sporleder, C, van den Bosch, A & Zervanou, K** eds *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bingenheimer, M, Hung, J-J & Wiles, S 2011 Social network visualization from TEI data. *Literary and Linguistic Computing* 26 (3) 271-78.

Bischof, F X & Maurer, H 2016 Konstanz (Diözese). *Historical Dictionary of Switzerland (HDS).* http://www.hls-dhs-dss.ch/textes/d/D7016.php (version date 2016/04/13, accessed September 2016).

Blei, D M, Ng, A Y & Jordan, M I 2003 Latent dirichlet allocation. *Journal of Machine Learning Research* 3 993-1022.

Blondel, V D, Guillaume, J-L, Lambiotte, R & Lefebvre, E 2008 Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008 P10008.

Bodenhammer, D J, Corrigan, J & Harris, T M 2010 *The Spatial Humanities - GIS and the Future of Humanities Scholarship* Indiana University Press, Bloomington, IN, USA.

Bodenhammer, D J, Harris, T M & Corrigan, J 2013 Deep Mapping and the Spatial Humanities. *International Journal of Humanities and Arts Computing* 7 170-75.

Boren, M T & Ramey, J 2000 Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication* 43 (3) 261-78.

Borlund, P 2009 User-Centred Evaluation of Information Retrieval Systems. in **Göker, A & Davies, J** eds *Information Retrieval - Searching in the 21st Century.* John Wiley & Sons, Ltd, Chichester, UK.

Bradley, J V 1960 Chapter 8 - Run of constant probability events. *Distribution-free statistical tests.* WADD Technical Report, Behavioral Sciences Laboratory Aerospace Medical Division, Ohio, OH, USA 195-221.

Branke, J 2001 Dynamic Graph Drawing. in **Kaufmann, M & Wagner, D** eds *Drawing Graphs: Methods and Models.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Brassel-Moser, R 2004 Crossair. *Historical Dictionary of Switzerland (HDS).* http://www.hls-dhs-dss.ch/textes/d/D41832.php (version date 2004/03/11, accessed August 2016).

Brin, S & Page, L 1998 The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30 107-17.

Brooke, J 1996 SUS - A 'quick and dirty' usability scale. in **Jordan, P W, Thomas, B, Weerdmeester, B A & McClelland, I L** eds *Usability evaluation in industry.* Taylor & Francis Ltd, London, UK.

Bruggmann, A & Fabrikant, S I 2014 Spatializing time in a history text corpus. in **Stewart, K, Pebesma, E, Navratil, G, Fogliaroni, P & Duckham, M** eds *Proceedings of the 8th International Conference on Geographic Information Science (GIScience 2014).* Vienna, Austria 183-86.

Bruggmann, A & Fabrikant, S I 2016 How does GIScience support spatio-temporal information search in the humanities? *Spatial Cognition & Computation* 1-17.

Bruggmann, A, Salvini, M M & Fabrikant, S I 2013 Cartograms of self-organizing maps to explore user-generated content. *Proceedings of the 26th International Cartographic Conference (ICC).* Dresden, Germany.

Brunsdon, C, Fotheringham, A S & Charlton, M E 1996 Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 28 (4) 281-98.

Bucher, B, Clough, P, Joho, H, Purves, R & Syed, A K 2005 Geographic IR systems: requirements and evaluation. *Proceedings of the 22nd International Cartographic Conference (ICC).* A Coruña, Spain.

Buscaldi, D 2011 Approaches to Disambiguating Toponyms. in **Tanin, E** ed *The SIGSPATIAL Special - Letters on Geographic Information Retrieval.* 3 (2) 16-19.

Byrt, T, Bishop, J & Carlin, J B 1993 Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46 (5) 423-29.

Börner, K, Chen, C & Boyack, K W 2003 Visualizing knowledge domains. *Annual Review of Information Science and Technology* 37 (1) 179-255.

Card, S K 1996 Visualizing retrieved information: a survey. *IEEE Computer Graphics and Applications* 16 (2) 63-67.

Card, S K, Mackinlay, J D & Shneiderman, B 1999 *Readings in information visualization: using vision to think* Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA.

Card, S K, Robertson, G G & Mackinlay, J D 1991 The information visualizer, an information workspace. in **Robertson, S P, Olson, G M & Olson, J S** eds *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New Orleans, LA, USA.

Castells, M 2010 *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I, 2nd Edition with a New Preface* Wiley-Blackwell.

Chang, J, Boyd-Graber, J, Gerrish, S, Wang, C & Blei, D M 2009 Reading Tea Leaves: How Humans Interpret Topic Models. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems.* Vancouver, Canada.

Cisco 2015 The Zettabyte Era: Trends and Analysis. Cisco Systems, Inc., http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html (accessed March 2016).

Ciula, A, Spence, P & Vieira, J M 2008 Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project. *Literary and Linguistic Computing* 23 (3) 311-25.

Clifford, J, Alex, B, Coates, C M, Klein, E & Watson, A 2016 Geoparsing history: Locating commodities in ten million pages of nineteenth-century sources. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49 115-31.

Cohen, J 1960 A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20 (1) 37-46.

Couclelis, H 1998 Worlds of Information: The Geographic Metaphor in the Visualization of Complex Information. *Cartography and Geographic Information Systems* 25 (4) 209-20.

Couclelis, H 2005 Space, time, geography. in **Longley, P A, Goodchild, M F, Maguire, D J & Rhind, D W** eds *Geographical Information Systems: Principles, Techniques, Management and Applications, 2nd Edition, Abridged.* John Wiley & Sons, Ltd, Chichester, UK.

Cöltekin, A, Heil, B, Garlandini, S & Fabrikant, S I 2009 Evaluating the Effectiveness of Interactive Map Interface Designs: A Case Study Integrating Usability Metrics with Eye-Movement Analysis. *Cartography and Geographic Information Science* 36 (1) 5-17.

Dear, M 2015 Practicing Geohumanities. *GeoHumanities* 1 (1) 20-35.

Dearholt, D W & Schvaneveldt, R W 1990 Properties of pathfinder networks. in **Schvaneveldt, R W** ed *Pathfinder associative networks.* Ablex Publishing Corp., Norwood, NJ, USA 1-30.

Deerwester, S, Dumais, S T, Furnas, G W, Landauer, T K & Harshman, R 1990 Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 391-407.

Derczynski, L, Strötgen, J, Campos, R & Alonso, O 2015 Time and information retrieval: Introduction to the special issue. *Information Processing and Management* 51 (6) 786-90.

Derungs, C 2014 *From Text to Landscape: Extraction of Landscape Concepts through the Resolution of Ambiguity and Vagueness present in Descriptions of Natural Landscapes* Department of Geography, University of Zurich, Zurich, Switzerland.

Derungs, C & Purves, R S 2014 From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science* 28 (6) 1272-93.

DeWalt, K M & DeWalt, B R 2002 *Participant Observation: A Guide for Fieldworkers* Rowman & Littlefield Pub Incorporated, Lanham, MD, USA.

DiBiase, D, MacEachren, A M, Krygier, J B & Reeves, C 1992 Animation and the Role of Map Design in Scientific Visualization. *Cartography and Geographic Information Systems* 19 (4) 201-14.

Domo 2015 Data Never Sleeps 3.0. Domo, Inc., https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/ (accessed March 2016).

Donaldson, C, Gregory, I N & Taylor, J E 2016 Implementing Corpus Analysis and GIS to Examine Historical Accounts of the English Lake District. *The Making of a Historical Atlas*. Northeast Asian History Foundation (in press).

Drucker, J 2011a Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 5 (1).

Drucker, J 2011b Humanities Approaches to Interface Theory. *Culture Machine* 12. http://www.culturemachine.net/index.php/cm/article/view/434 (accessed April 2016).

Dubin, D 2004 The Most Influential Paper Gerard Salton Never Wrote. *Library Trends* 52 (4) 748-64.

Dumais, S T 2004 Latent semantic analysis. *Annual Review of Information Science and Technology* 38 (1) 188-230.

Ellis, D, Furner-Hines, J & Willett, P 1993 Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management* 3 (2) 128-49.

Evans, C & Jasnow, B 2014 Mapping Homer's Catalogue of Ships. *Literary and Linguistic Computing* 29 (3) 317-25.

Fabrikant, S I 2000 Spatialized Browsing in Large Data Archives. *Transactions in GIS* 4 (1) 65-78.

Fabrikant, S I & Buttenfield, B P 2001 Formalizing Semantic Spaces For Information Access. *Annals of the Association of American Geographers* 91 (2) 263-80.

Fabrikant, S I, Gabathuler, C & Skupin, A 2015 SOMViz: Web-based Self-Organizing Maps. *Kartographische Nachrichten - Journal of Cartography and Geographic Information* 65 (2) 81-91.

Fabrikant, S I, Montello, D R & Mark, D M 2006 The distance similarity metaphor in region-display spatializations. *IEEE Computer Graphics and Applications* 26 (4) 34-44.

Fabrikant, S I, Montello, D R, Ruocco, M & Middleton, R S 2004 The distance–similarity metaphor in network-display spatializations. *Cartography and Geographic Information Science* 31 (4) 237-52.

Fabrikant, S I & Salvini, M M 2011 Charting the ICA World of Cartography 1999-2009. *Proceedings of the 25th International Cartographic Conference (ICC)*. Paris, France.

Fabrikant, S I & Skupin, A 2005 Cognitively Plausible Information Visualization. in **Dykes, J, MacEachren, A M & Kraak, M-J** eds *Exploring Geovisualization*. Elsevier Ltd, Amsterdam, Netherlands 667-90.

Fayyad, U, Piatetsky-Shapiro, G & Smyth, P 1996 From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 17 (3) 37-54.

Feinstein, A R & Cicchetti, D V 1990 High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology* 43 (6) 543-49.

Ferro, L, Gerber, L, Mani, I, Sundheim, B & Wilson, G 2005 TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical Report, The MITRE Corporation.

Ferro, L, Mani, I, Sundheim, B & Wilson, G 2001 TIDES Temporal Annotation Guidelines, Version 1.0.2. Technical Report, The MITRE Corporation.

Fink, U 2011 Schweizerische Kirchenzeitung. *Historical Dictionary of Switzerland (HDS).* http://www.hls-dhs-dss.ch/textes/d/D49647.php (version date 2011/10/27, accessed August 2016).

Fitzpatrick, K 2012 The Humanities, Done Digitally. in **Gold, M K** ed *Debates in the Digital Humanities.* University of Minnesota Press, Minneapolis, MN, USA 12-15.

Fotheringham, A S 2009 Geographically Weighted Regression. in **Fotheringham, A S & Rogerson, P A** eds *The SAGE handbook of Spatial Analysis.* SAGE Publications Ltd, London, UK.

Fotheringham, A S, Brunsdon, C & Charlton, M 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* John Wiley & Sons, Ltd, Chichester, UK.

Frick, A, Ludwig, A & Mehldau, H 1995 A fast adaptive layout algorithm for undirected graphs (extended abstract and system demonstration). in **Tamassia, R & Tollis, I G** eds *Graph Drawing: DIMACS International Workshop, GD '94 Princeton, New Jersey, USA, October 10–12, 1994 Proceedings.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Friedl, J E F 2006 *Mastering Regular Expressions, 3rd edition* O'Reilly Media.

Gan, Q, Attenberg, J, Markowetz, A & Suel, T 2008 Analysis of geographic queries in a search engine log. *Proceedings of the 1st international workshop on Location and the Web (LocWeb 2008).* Beijing, China 49-56.

Geary, R C 1954 The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician* 5 115-46.

Getis, A & Ord, J K 1992 The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24 (3) 189-206.

Gibbs, F & Owens, T 2012 Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly* 6 (2).

Gisev, N, Bell, J S & Chen, T F 2013 Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy* 9 (3) 330-38.

Gold, M K 2012 The Digital Humanities Moment. in **Gold, M K** ed *Debates in the Digital Humanities.* University of Minnesota Press, Minneapolis, MN, USA.

Golub, G H & Reinsch, C 1970 Singular value decomposition and least squares solutions. *Numerische Mathematik* 14 (5) 403-20.

Goodchild, M F 2002 Finding the mainstream. *Proceedings of the E-future : into the mainstream, a joint AURISA and Institution of Surveyors Australia conference.* Adelaide, Australia.

Gooding, P 2013 Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing* 28 (3) 425-31.

Graham, M & De Sabbata, S 2015 Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A* 47 (6) 1254-64.

Gregory, I 2010 Exploiting Time and Space: A challenge for GIS in the Digital Humanities. in **Bodenhammer, D J, Corrigan, J & Harris, T M** eds *The*

Spatial Humanities - GIS and the Future of Humanities Scholarship. Bloomington, IN, USA 58-75.

Gregory, I, Atkinson, P, Hardie, A, Joulain-Jay, A, Kershaw, D, Porter, C, Rayson, P & Rupp, C J 2016 From digital resources to historical scholarship with the British Library 19th Century Newspaper Collection. *Journal of Siberian Federal University: Humanities and social sciences* 9 (4) 994-1006.

Gregory, I N & Ell, P S 2007 *Historical GIS - Technologies, Methodologies, and Scholarschip* Cambridge University Press, Cambridge, UK.

Gregory, I N & Hardie, A 2011 Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistic Computing* 26 (3) 297-314.

Gregory, I N & Healey, R G 2007 Historical GIS: structuring, mapping and analysing geographies of the past. *Progress in Human Geography* 31 (5) 638-53.

Griffiths, T L & Steyvers, M 2004 Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 (Supplement 1) 5228-35.

Grinstein, G, Kobsa, A, Plaisant, C & Stasko, J T 2003 Which comes first, usability or utility? *Proceedings of the 14th IEEE Visualization Conference (VIS'03)*. Seattle, WA, USA 605-06.

Grossner, K 2014 Geographic Search. in **Ballatore, A, Hegarty, M, Kuhn, W & Parsons, E** eds *Position Papers, 2014 Specialist Meeting - Spatial Search*. Santa Barbara, CA, USA 26-28.

Grover, C, Tobin, R, Byrne, K, Woollard, M, Reid, J, Dunn, S & Ball, J 2010 Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences* 368 3875-89.

Haber, P 2007 Historisches Lexikon der Schweiz. *Traverse* 14 (1) 127-33.

Haber, P 2008 Die Vision eines e-HLS der Zukunft. in **Jorio, M & Eggs, C** eds *Am Anfang ist das Wort - Lexika in der Schweiz*. Hier und Jetzt, Baden, Switzerland 135-47.

Hacioglu, K, Chen, Y & Douglas, B 2005 Automatic Time Expression Labeling for English and Chinese Text. in **Gelbukh, A** ed *Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg 548-59.

Hadlak, S, Schumann, H & Schulz, H-J 2015 A Survey of Multi-faceted Graph Visualization. *Proceedings of the Eurographics Conference on Visualization (EuroVis)*. Cagliari, Italy.

Hagenauer, J & Helbich, M 2013 Hierarchical self-organizing maps for clustering spatiotemporal data. *International Journal of Geographical Information Science* 27 (10) 2026-42.

Hahmann, S & Burghardt, D 2013 How much information is geospatially referenced? Networks and cognition. *International Journal of Geographical Information Science* 27 (6) 1171-89.

Hahmann, S, Burghardt, D & Weber, B 2011 "80% of All Information is Geospatially Referenced"??? Towards a Research Framework: Using the Semantic Web for (In) Validating this Famous Geo Assertion. in **Geertman, S, Reinhardt, W & Toppen, F** eds *Proceedings of the 14th AGILE International Conference on Geographic Information Science*. Utrecht, Netherlands.

Haining, R 2009 The Special Nature of Spatial Data. in **Fotheringham, A S & Rogerson, P A** eds *The SAGE handbook of Spatial Analysis*. SAGE Publications Ltd, London, UK.

Hawkins, H, Cabeen, L, Callard, F, Castree, N, Daniels, S, DeLyser, D, Neely, H M & Mitchell, P 2015 What Might GeoHumanities Do? Possibilities, Practices, Publics, and Politics. *GeoHumanities* 1 (2) 211-32.

Hecht, B & Raubal, M 2008 GeoSR: Geographically Explore Semantic Relations in World Knowledge. in **Bernard, L, Friis-Christensen, A & Pundt, H** eds *Proceedings of the 11th AGILE International Conference on Geographic Information Science.* Girona, Spain 95-113.

Hiemstra, D 1998 A Linguistically Motivated Probabilistic Model of Information Retrieval. *Research and Advanced Technology for Digital Libraries: Second European Conference, ECDL'98 Heraklion, Crete, Greece September 21–23, 1998 Proceedings.* Springer Berlin Heidelberg, Berlin, Heidelberg 569-84.

Hill, L L 2006 *Georeferencing - The geographic associations of information* MIT Press, Cambridge, MA, USA; London, UK.

Hill, L L, Frew, J & Zheng, Q 1999 Geographic Names - The implementation of a Gazetteer in a Georeferenced Digital Library. *D-Lib Magazine* 5 (1).

Hinrichs, U, Alex, B, Clifford, J, Watson, A, Quigley, A, Klein, E & Coates, C M 2015 Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities* 30 (Supplement 1) i50-i75.

Historical Dictionary of Switzerland (HDS) 2016a, http://www.hls-dhs-dss.ch/d/home (accessed June 2016).

Historical Dictionary of Switzerland (HDS) 2016b e-HDS search interface (German version). http://www.hls-dhs-dss.ch/d/home (accessed June 2016).

Historical Dictionary of Switzerland (HDS) 2016c Management summary New HDS (German version). http://www.hls-dhs-dss.ch/redac/downloads/summaryD.pdf (accessed June 2016).

Hofmann, T 1999 Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, Berkeley, CA, USA 50-57.

Horisberger, B, Huber, A, Illi, M, König, M, Suter, M & Windler, R 2015 Zürich (Kanton). *Historical Dictionary of Switzerland (HDS).* http://www.hls-dhs-dss.ch/textes/d/D7381.php (version date 2015/02/03, accessed June 2016).

Hossain, M A, Dwivedi, Y K & Rana, N P 2016 State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of Organizational Computing and Electronic Commerce* 26 (1-2) 14-40.

Hubler, L 2004 Ennetbirgische Vogteien. *Historical Dictionary of Switzerland (HDS).* Translated from French. http://www.hls-dhs-dss.ch/textes/d/D46979.php (version date 2004/11/01, accessed August 2016).

Hörsch, W 2008 Landvogt [Obervogt, Vogt]. *Historical Dictionary of Switzerland (HDS).* http://www.hls-dhs-dss.ch/textes/d/D26435.php (version date 2008/11/13, accessed August 2016).

Ingwersen, P & Järvelin, K 2005 *The Turn - Integration of Information Seeking and Retrieval in Context* Springer Netherlands.

ITU 2015 Measuring the Information Society Report 2015. International Telecommunication Union (ITU), http://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2015/MISR2015-w5.pdf (accessed March 2016).

Janetzko, H, Jäckle, D & Schreck, T 2014 Geo-Temporal Visual Analysis of Customer Feedback Data Based on Self-Organizing Sentiment Maps. *International Journal on Advances in Intelligent Systems* 7 (1-2).

Jeffers, J N R 1973 A Basic Subroutine for Geary's Contiguity Ratio. *Journal of the Royal Statistical Society. Series D (The Statistician)* 22 (4) 299-302.

Jenks, G 1967 The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography* 7 186-90.

Jockers, M L 2013 *Macroanalysis: Digital Methods & Literary History* University of Illinois Press, Urbana, IL, USA.

Jolliffe, I T 2002 *Principle Component Analysis* Springer, New York, NY, USA.

Jones, C B & Purves, R S 2008 Geographical information retrieval. *International Journal of Geographical Information Science* 22 (3) 219-28.

Jones, C B & Purves, R S 2009 Geographical information retrieval. in **Liu, L & Özsu, M T** eds *Encyclopedia of Database Systems, Part 7*. Springer, New York, NY, USA 1227-31.

Jones, R, Zhang, W V, Rey, B, Jhala, P & Stipp, E 2008 Geographic intention and modification in web search. *International Journal of Geographical Information Science* 22 (3) 229-46.

Jorio, M 1998 Das neue Historische Lexikon der Schweiz (HLS). *Informationsmittel für Bibliotheken* 6 (1/2-169).

Jorio, M 2000 Das Historische Lexikon der Schweiz im Jahre 2000. *Schweizerische Zeitschrift für Geschichte - Revue suisse d'histoire - Rivista storica svizzera* 50 (2) 198-203.

Jorio, M 2004 Die Geschichte der Enzyklopädie in der Schweiz seit dem 17. Jahrhundert. in **Stammen, T & Weber, W E J** eds *Wissenssicherung, Wissensordnung und Wissensverarbeitung - Das europäische Modell der Enzyklopädien*. Akademie Verlag, Berlin, Germany 105-17.

Jorio, M 2014, September 13 Enzyklopädisches Wissen im digitalen Zeitalter - Das "Historische Lexikon der Schweiz" und die Medienrevolution. *Neue Zürcher Zeitung*. 61.

Jänicke, S, Franzini, G, Cheema, M F & Scheuermann, G 2015 On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *Proceedings of the Eurographics Conference on Visualization (EuroVis)*. Cagliari, Italy.

Kaplan, F 2015 A map for Big Data research in Digital Humanities. *Frontiers in Digital Humanities* 2.

Keim, D A 2002 Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8 (1) 1-8.

Keim, D A, Krstajic, M, Rohrdantz, C & Schreck, T 2013 Real-Time Visual Analytics for Text Streams. *Computer* 46 (7) 47-55.

Keim, D A, Mansmann, F, Schneidewind, J & Ziegler, H 2006 Challenges in Visual Data Analysis. *Proceedings of the 10th International Conference on Information Visualisation (IV'06)*. London, UK.

Kelly, D 2009 Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3 (2) 1-224.

Kessler, F 2000 Focus Groups as a Means of Qualitatively Assessing the U-Boat Narrative. *Cartographica: The International Journal for Geographic Information and Geovisualization* 37 (4) 33-60.

Kirschenbaum, M 2012 What Is Digital Humanities and What's It Doing in English Departments? in **Gold, M K** ed *Debates in the Digital Humanities*. University of Minnesota Press, Minneapolis, MN, USA 3-11.

Kirschenbaum, M G 2004 "So the Colors Cover the Wires": Interface, Aesthetics, and Usability. in **Schreibman, S, Siemens, R & Unsworth, J** eds *A Companion to Digital Humanities*. Blackwell Publishing Ltd, Malden, MA, USA.

Kohler, F 2013 Pruntrut (Gemeinde). *Historical Dictionary of Switzerland (HDS).* Translated from French. http://www.hls-dhs-dss.ch/textes/d/D3003.php (version date 2013/08/08, accessed August 2016).

Kohonen, T 1990 The self-organizing map. *Proceedings of the IEEE* 78 (9) 1464-80.

Kohonen, T 2001 *Self-organizing maps* Springer Berlin Heidelberg, Berlin, Heidelberg.

Kos, A J & Psenicka, C 2000 Measuring Cluster Similarity across Methods. *Psychological Reports* 86 (3) 858-62.

Koua, E L & Kraak, M-J 2005 Evaluating Self-organizing Maps for Geovisualization. in **Dykes, J, MacEachren, A M & Kraak, M-J** eds *Exploring Geovisualization.* Elsevier Ltd, Amsterdam, Netherlands 627-43.

Kraak, M-J 2008 Editorial From Geovisualisation Toward Geovisual Analytics. *The Cartographic Journal* 45 (3) 163-64.

Kraak, M-J & MacEachren, A M 2005 Geovisualization and GIScience. *Cartography and Geographic Information Science* 32 (2) 67-68.

Kuhn, W 1992 Paradigms of GIS Use. *Proceedings of the 5th International Symposium on Spatial Data Handling.* Charleston, SC, USA.

Kuhn, W 1996 Handling Data Spatially: Spatializing User Interfaces. in **Kraak, M-J & Molenaar, M** eds *Proceedings of the 7th International Symposium on Spatial Data Handling.* Delft, Netherlands.

Kuhn, W & Blumenthal, B 1996 *Spatialization: Spatial Metaphors for User Interfaces* Technical University Vienna, Vienna, Austria.

Kuhn, W & Frank, A U 1991 A Formalization of Metaphors and Image-Schemas in User Interfaces. in **Mark, D M & Frank, A U** eds *Cognitive and Linguistic Aspects of Geographic Space.* Springer Netherlands, Dordrecht, Netherlands.

Lacayo-Emery, M A 2011 *An Integrated Toolset for Exploration of Spatio-Temporal Data Using Self-Organizing Maps and GIS* San Diego State University, San Diego, CA, USA.

Lakoff, G 1987 *Women, Fire, and Dangerous Things - What Categories Reveal about the Mind* The University of Chicago Press, Chicago, IL, USA and London, UK.

Lakoff, G & Johnson, M 1980 *Metaphors we live by* The University of Chicago Press, Chicago, IL, USA and London, UK.

Landauer, T K 1995 *The Trouble with Computers - Usefulness, Usability, and Productivity* MIT Press, Cambridge, MA, USA; London, UK.

Landauer, T K & Dumais, S T 1997 A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104 (2) 211-40.

Landay, J A & Myers, B A 2001 Sketching interfaces: toward more human interface design. *Computer* 34 (3) 56-64.

Landis, J R & Koch, G G 1977 The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33 (1) 159-74.

Larson, R R 1996 Geographic Information Retrieval and Spatial Browsing. in **Smith, L C & Gluck, M** eds *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information.* University of Illinois at Urbana-Champaign 81-124.

Laube, P & Purves, R S 2011 How fast is a cow? Cross-Scale Analysis of Movement Data. *Transactions in GIS* 15 401–18.

Le Tensorer, J-M 2015 Paläolithikum. *Historical Dictionary of Switzerland (HDS).* Translated from French. http://www.hls-dhs-dss.ch/textes/d/D8010.php (version date 2015/06/29, accessed June 2016).

Leidner, J L 2007 Toponym Resolution in Text. *School of Informatics, Institute for Communicating and Collaborative Systems.* University of Edinburgh, Edinburgh, UK.

Leidner, J L & Lieberman, M D 2011 Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. in **Tanin, E** ed *The SIGSPATIAL Special - Letters on Geographic Information Retrieval.* 3 (2) 5-11.

Leidner, J L, Sinclair, G & Webber, B 2003 Grounding spatial named entities for information extraction and question answering. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references.* Edmonton, Canada.

Leung, Y, Mei, C-L & Zhang, W-X 2000 Statistical Tests for Spatial Nonstationarity Based on the Geographically Weighted Regression Model. *Environment and Planning A* 32 (1) 9-32.

Leveling, J 2011 Challenges for Indexing in GIR. in **Tanin, E** ed *The SIGSPATIAL Special - Letters on Geographic Information Retrieval.* 3 (2) 29-32.

Leveling, J & Hartrumpf, S 2008 On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science* 22 (3) 289-99.

Leveling, J & Veiel, D 2007 Experiments on the Exclusion of Metonymic Location Names from GIR. in **Peters, C, Clough, P, Gey, F C, Karlgren, J, Magnini, B, Oard, D W, Rijke, M & Stempfhuber, M** eds *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.* Springer Berlin Heidelberg, Berlin, Heidelberg 901-04.

Lewis, C & Rieman, J 1993 *Task-centered user interface design: A practical introduction* University of Colorado, Boulder, Boulder, CO, USA.

Lewis, C & Wharton, C 1997 Chapter 30 - Cognitive Walkthroughs. in **Helander, M G, Landauer, T K & Prabhu, P V** eds *Handbook of Human-Computer Interaction, 2nd Edition.* North-Holland, Amsterdam, Netherlands.

Li, H, Srihari, R K, Niu, C & Li, W 2002 Location normalization for information extraction. *Proceedings of the 19th international conference on Computational Linguistics.* Association for Computational Linguistics, Taipei, Taiwan.

Li, H, Srihari, R K, Niu, C & Li, W 2003 InfoXtract location normalization: a hybrid approach to geographic references in information extraction. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references.* Association for Computational Linguistics, Edmonton, Canada.

Light, R J 1971 Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin* 76 (5) 365-77.

Loebbecke, C & Picot, A 2015 Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems* 24 (3) 149-57.

Luo, W, Yin, P, Di, Q, Hardisty, F & MacEachren, A M 2014 A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network. *PLoS ONE* 9 (2) e88666.

MacEachren, A M 1995 How maps work: Representation, Visualization, and Design. The Guilford Press, New York, NY, USA.

MacEachren, A M & Kraak, M-J 2001 Research Challenges in Geovisualization. *Cartography and Geographic Information Science* 28 (1) 3-12.

MacQueen, J 1967 Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics.* University of California Press, Berkeley, CA, USA 281-97.

Mahatody, T, Sagar, M & Kolski, C 2010 State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions. *International Journal of Human-Computer Interaction* 26 (8) 741-85.

Mandl, T 2011 Evaluating GIR: geography-oriented or user-oriented? in **Tanin, E** ed *The SIGSPATIAL Special - Letters on Geographic Information Retrieval.* 3 (2) 42-45.

Manning, C D, Raghavan, P & Schütze, H 2009a Boolean retrieval. *Introduction to Information Retrieval - Online edition.* Cambridge, UK. http://nlp.stanford.edu/IR-book/pdf/01bool.pdf (accessed September 2016).

Manning, C D, Raghavan, P & Schütze, H 2009b Evaluation in information retrieval. *Introduction to Information Retrieval - Online edition.* Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/pdf/08eval.pdf (accessed May 2016).

Manning, C D, Raghavan, P & Schütze, H 2009c Probabilistic information retrieval. *Introduction to Information Retrieval - Online edition.* Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/pdf/11prob.pdf (accessed July 2016).

Manning, C D, Raghavan, P & Schütze, H 2009d Scoring, term weighting and the vector space model. *Introduction to Information Retrieval - Online edition.* Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/pdf/06vect.pdf (accessed April 2016).

Manning, C D, Raghavan, P & Schütze, H 2009e The term vocabulary and postings lists. *Introduction to Information Retrieval - Online edition.* Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/pdf/02voc.pdf (accessed June 2016).

Martins, B, Anastácio, I & Calado, P 2010 A Machine Learning Approach for Resolving Place References in Text. in **Painho, M, Santos, Y M & Pundt, H** eds *Geospatial Thinking.* Springer Berlin Heidelberg, Berlin, Heidelberg 221-36.

McCallum, A K 2002 MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu/ (accessed June 2016).

Mikheev, A, Moens, M & Grover, C 1999 Named Entity recognition without gazetteers. *Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics.* Bergen, Norway.

Mimno, D 2012 Computational Historiography: Data Mining in a Century of Classics Journals. *Journal on Computing and Cultural Heritage* 5 (1) 3:1-3:19.

Mimno, D, Wallach, H W, Talley, E, Leenders, M & McCallum, A 2011 Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011).* Edinburgh, UK.

Montello, D R, Fabrikant, S I, Ruocco, M & Middleton, R S 2003 Testing the First Law of Cognitive Geography on Point-Display Spatializations. in **Kuhn, W, Worboys, M F & Timpf, S** eds *Spatial Information Theory. Foundations of Geographic Information Science: International Conference, COSIT 2003, Kartause Ittingen, Switzerland, September 24-28, 2003. Proceedings.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Moran, P A P 1950 Notes on Continuous Stochastic Phenomena. *Biometrika* 37 (1/2) 17-23.

Moretti, F 2005 *Graphs, Maps, Trees: Abstract Models for Literary History* Verso, London, UK.

Moretti, F 2013 *Distant Reading* Verso, London, UK.

Morosoli, R 2000 Blickpunkt: Historisches Lexikon der Schweiz (HLS), Arbeitsstelle Kanton Zug. *Tugium* 16 9-16.

Muelder, C, Gou, L, Ma, K-L & Zhou, M X 2014 Multivariate Social Network Visual Analytics. in **Kerren, A, Purchase, H C & Ward, M O** eds *Multivariate Network Visualization: Dagstuhl Seminar #13201, Dagstuhl Castle, Germany, May 12-17, 2013, Revised Discussions.* Springer International Publishing, Cham, Switzerland 37-59.

Murchú, T Ó & Lawless, S 2014 The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data. *Proceedings of the Digital Humanities 2014 conference.* Lausanne, Switzerland.

Murrieta-Flores, P, Baron, A, Gregory, I, Hardie, A & Rayson, P 2015 Automatically Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in the 19th Century. *Transactions in GIS* 19 (2) 296-320.

Nelson, K J, Quinn, S, Swedberg, B, Chu, W & MacEachren, M A 2015 Geovisual Analytics Approach to Exploring Public Political Discourse on Twitter. *ISPRS International Journal of Geo-Information* 4 337-66.

Nielsen, J 1994 Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies* 41 (3) 385-97.

Nielsen, J 2012 How Many Test Users in a Usability Study? Nielsen Norman Group, https://www.nngroup.com/articles/how-many-test-users/ (accessed September 2016).

North, C 2006 Toward measuring visualization insight. *IEEE Computer Graphics and Applications* 26 (3) 6-9.

NWB Team 2006 Network Workbench Tool. 1.0.0 ed. Indiana University, Northeastern University and University of Michigan, http://nwb.cns.iu.edu/ (accessed July 2016).

Okabe, A, Boots, B, Sugihara, K, Chiu, S N & Kendall, D G 2000 *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd Edition* John Wiley & Sons, Ltd, Chichester, UK.

Onwuegbuzie, A J, Dickinson, W B, Leech, N L & Zoran, A G 2009 A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research. *International Journal of Qualitative Methods* 8 (3) 1-21.

Openshaw, S 1983 The modifiable areal unit problem. *Concepts and techniques in modern geography* 38 1-41.

Palacio, D, Cabanac, G, Sallaberry, C & Hubert, G 2011 On the evaluation of Geographic Information Retrieval systems. *International Journal on Digital Libraries* 11 (2) 91-109.

Palacio, D, Derungs, C & Purves, R S 2015 Development and evaluation of a geographic information retrieval system using fine grained toponyms. *Journal of Spatial Information Science (JOSIS)* 11 1-29.

Park, T K 1994 Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American Society for Information Science* 45 (3) 135-41.

Peters, C & Braschler, M 2001 European research letter: Cross-language system evaluation: The CLEF campaigns. *Journal of the American Society for Information Science and Technology* 52 (12) 1067-72.

Peuquet, D J 2002 *Representations of Space and Time* The Guildford Press, New York, NY, USA.

Pfeffer, J, Mvrar, A & Batagelj, V 2013 txt2pajek: Creating Pajek Files from Text Files. Technical Report CMU-ISR-13-110, Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Ponte, J M & Croft, W B 1998 A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* Melbourne, Australia 275-81.

Porter, M F 1980 An algorithm for suffix stripping. *Program* 14 (3) 130-37.

Pouliquen, B, Kimler, M, Steinberger, R, Ignat, C, Oellinger, T, Blackler, K, Fuart, F, Zaghouani, W, Widiger, A, Forslund, A-C & Best, C 2006 Geocoding multilingual texts: Recognition, disambiguation and visualisation.

Proceedings of the 5th International Conference on Language Ressources and Evaluation. Genoa, Italy.

Purchase, H C & Samra, A 2008 Extremes Are Better: Investigating Mental Map Preservation in Dynamic Graphs. in **Stapleton, G, Howse, J & Lee, J** eds *Diagrammatic Representation and Inference: 5th International Conference, Diagrams 2008, Herrsching, Germany, September 19-21, 2008. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Purves, R S, Clough, P, Jones, C B, Arampatzis, A, Bucher, B, Finch, D, Fu, G, Joho, H, Syed, A K, Vaid, S & Yang, B 2007 The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science* 21 (7) 717-45.

Pustejovsky, J, Knippen, R, Littman, J & Saurí, R 2005 Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation* 39 (2) 123-64.

Radlinski, F & Joachims, T 2005 Query chains: learning to rank from implicit feedback. *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*. Chicago, IL, USA 239-48.

Ramsay, S 2011 Who's In and Who's Out. in **Terras, M, Nyhan, J & Vanhoutte, E** eds *Defining Digital Humanities - A Reader*. Ashgate Publishing Ltd, Farnham, UK 239-42.

Rauch, E, Bukatin, M & Baker, K 2003 A confidence-based framework for disambiguating geographic terms. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*. Association for Computational Linguistics, Edmonton, Canada 50-54.

Reiter, N, Frank, A & Hellwig, O 2014 An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing* 29 (4) 583-605.

Reitsma, R & Trubin, S 2007 Information space partitioning using adaptive Voronoi diagrams. *Information Visualization* 6 123-38.

Robertson, G G, Mackinlay, J D & Card, S K 1991 Cone Trees: animated 3D visualizations of hierarchical information. in **Robertson, S P, Olson, G M & Olson, J S** eds *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New Orleans, LA, USA.

Robertson, S & Zaragoza, H 2009 The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3 (4) 333-89.

Robinson, A C, Peuquet, D J, Pezanowski, S, Hardisty, F A & Swedberg, B 2016 Design and evaluation of a geovisual analytics system for uncovering patterns in spatio-temporal event data. *Cartography and Geographic Information Science* 1-13.

Rosson, M B & Carroll, J M 2002a Chapter 2 - Analyzing Requirements. *Usability Engineering - Scenario-Based Development of Human-Computer Interaction*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA 37-78.

Rosson, M B & Carroll, J M 2002b *Usability Engineering - Scenario-Based Development of Human-Computer Interaction* Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA.

Roth, R E & MacEachren, A M 2016 Geovisual analytics and the science of interaction: an empirical interaction study. *Cartography and Geographic Information Science* 43 (1) 30-54.

Roth, R E, MacEachren, A M, Andrienko, G, Andrienko, N, Dykes, J, Kraak, M-J, Robinson, A C & Schumann, H 2014 Geovisual Analytics & the Science of Interaction: A Case Study. *Proceedings of the workshop on GeoVisual Analytics: Interactivity, Dynamics, and Scale in conjucntion with GIScience 2014*. Vienna, Austria.

Roth, R E, Ross, K S & MacEachren, A M 2015 User-Centered Design for Interactive Maps: A Case Study in Crime Analysis. *ISPRS International Journal of Geo-Information* 4 262-301.

Rubin, J & Chisnell, D 2008 *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, 2nd Edition* Wiley Publishing, Inc., Indianapolis, IN, USA.

Saffrey, P & Purchase, H 2008 The "mental map" versus "static aesthetic" compromise in dynamic graphs: a user study. *Proceedings of the 9th conference on Australasian user interface (AUIC '08).* Wollongong, NSW, Australia 85-93.

Salvini, M M 2012 *Spatialization von nutzergenerierten Inhalten für die explorative Analyse des globalen Städtenetzes* Department of Geography, University of Zurich, Zurich, Switzerland.

Salvini, M M & Fabrikant, S I 2016 Spatialization of user-generated content to uncover the multirelational world city network. *Environment and Planning B: Planning and Design* 43 228-48.

Sanderson, M 2010 Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4 (4) 247-375.

Sanderson, M & Croft, W B 2012 The History of Information Retrieval Research. *Proceedings of the IEEE* 100 1444-51.

Sanderson, M & Kohler, J 2004 Analyzing geographic queries. *Proceedings of the ACM SIGIR 2004 Workshop on Geographic Information Retrieval.* Sheffield, UK.

Santos, J, Anastácio, I & Martins, B 2014 Using machine learning methods for disambiguating place references in textual documents. *GeoJournal* 80 (3) 375-92.

Schilder, F & Habel, C 2001 From temporal expressions to temporal information: semantic tagging of news messages. *Proceedings of the workshop on Temporal and spatial information processing.* Toulouse, France.

Schmid, H 1994 Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK.

Schmid, H 1995 Improvements In Part-of-Speech Tagging With an Application To German. *Proceedings of the EACL SIGDAT Workshop.* Dublin, Ireland.

Schreibman, S, Siemens, R & Unsworth, J 2004 *A Companion to Digital Humanities* Blackwell Publishing Ltd, Oxford, UK.

Senn, H 2010 Militärorganisationen (MO). *Historical Dictionary of Switzerland (HDS).* http://www.hls-dhs-dss.ch/textes/d/D24630.php (version date 2010/05/21, accessed August 2016).

Shneiderman, B 1996 The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings of the IEEE Symposium on Visual Languages.* Boulder, CO, USA.

Shneiderman, B 2002 Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization* 1 (1) 5-12.

Simon, R, Barker, E, Isaksen, L & de Soto Cañamares, P 2015 Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito. *e-Perimetron* 10 (2) 49-59.

Skupin, A 2002 A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications* 22 (1) 50-58.

Skupin, A 2008 Spatialization. in **Kemp, K K** ed *Encyclopedia of Geographic Information Science.* SAGE Publications, Inc., Thousand Oaks, CA, USA 419-22.

Skupin, A 2009 Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization. *Journal of Informetrics* 3 (3) 233-45.

Skupin, A 2014 Making a Mark: a computational and visual analysis of one researcher's intellectual domain. *International Journal of Geographical Information Science* 28 (6) 1209-32.

Skupin, A & Agarwal, P 2008 Introduction: What is a Self-Organizing Map? *Self-Organising Maps*. John Wiley & Sons, Ltd.

Skupin, A, Biberstine, J R & Börner, K 2013 Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. *PLoS ONE* 8 (3) e58779.

Skupin, A & Buttenfield, B P 1996 Spatial metaphors for visualizing very large data archives. *Proceedings of the GIS/LIS conference*. Denver, CO, USA.

Skupin, A & Buttenfield, B P 1997 Spatial metaphors for visualizing information spaces. *Proceedings of the Auto-Carto 13 conference*. Seattle, WA, USA.

Skupin, A & de Jongh, C 2005 Visualizing the ICA - a content-based approach. *Proceedings of the 22nd International Cartographic Conference (ICC)*. A Coruña, Spain.

Skupin, A & Esperbé, A 2011 An alternative map of the United States based on an n-dimensional model of geographic space. *Journal of Visual Languages & Computing* 22 (4) 290-304.

Skupin, A & Fabrikant, S I 2003 Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science* 30 (2) 95-115.

Skupin, A & Fabrikant, S I 2007 Spatialization. in **Wilson, J P & Fotheringham, A S** eds *The Handbook of Geographic Information Science*. Blackwell Publishing Ltd, Oxford, UK 61-79.

Skupin, A & Hagelman, R 2005 Visualizing Demographic Trajectories with Self-Organizing Maps. *GeoInformatica* 9 (2) 159-79.

Smith, D A & Crane, G 2001 Disambiguating Geographic Names in a Historical Digital Library. in **Constantopoulos, P & Sølvberg, I T** eds *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001 Darmstadt, Germany, September 4-9, 2001 Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg 127-36.

Southall, H, Mostern, R & Berman, M L 2011 On historical gazetteers. *International Journal of Humanities and Arts Computing* 5 (2) 127-45.

Spiro, L 2012 "This Is Why We Fight": Defining the Values of the Digital Humanities. in **Gold, M K** ed *Debates in the Digital Humanities*. University of Minnesota Press, Minneapolis, MN, USA 16-35.

Steiger, E, Resch, B & Zipf, A 2016 Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science* 30 (9) 1694-716.

Stewart, G W 1993 On the early history of the singular value decomposition. *SIAM Review* 35 (4) 551-66.

Steyvers, M & Griffiths, T 2007 Probabilistic Topic Models. in **Landauer, T K, McNamara, D S, Dennis, S & Kintsch, W** eds *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ 427-48.

Steyvers, M & Griffiths, T L 2008 Rational analysis as a link between human memory and information retrieval. in **Chater, N & Oaksford, M** eds *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford University Press Inc., New York, NY, USA 329-49.

Strötgen, J 2015 *Domain-sensitive Temporal Tagging for Event-centric Information Retrieval* Department of Computer Science, Heidelberg University, Heidelberg, Germany.

Strötgen, J, Bögel, T, Zell, J, Armiti, A, Van Canh, T & Gertz, M 2014 Extending HeidelTime for Temporal Expressions Referring to Historic Dates. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland.

Strötgen, J & Gertz, M 2013 Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 47 (2) 269-98.

Svensson, P 2009 Humanities Computing as Digital Humanities. *Digital Humanities Quarterly* 3 (3).

Svensson, P 2012 Beyond the Big Tent. in **Gold, M K** ed *Debates in the Digital Humanities*. University of Minnesota Press, Minneapolis, MN, USA 36-49.

Terras, M 2011 Peering Inside the Big Tent. in **Terras, M, Nyhan, J & Vanhoutte, E** eds *Defining Digital Humanities - A Reader*. Ashgate Publishing Ltd, Farnham, UK 263-70.

Thomas, J J & Cook, K A 2005 *Illuminating the Path* IEEE Computer Society Press, Los Alamitos, CA, USA.

Thomas, J J & Cook, K A 2006 A visual analytics agenda. *IEEE Computer Graphics and Applications* 26 (1) 10-13.

Thompson, K 1968 Regular Expression Search Algorithm. *Communications of the ACM* 11 (6) 419-22.

Tinsley, H E A & Weiss, D J 1975 Interrater Reliability and Agreement of Subjective Judgments. *Journal of Counseling Psychology* 22 (4) 358-76.

Tobler, W R 1970 A computer movie simulating urban growth in the Detroit Region. *Economic Geography* 46 234-40.

Tomaszewski, B 2008 Producing Geo-historical Context from Implicit Sources: A Geovisual Analytics Approach. *The Cartographic Journal* 45 (3) 165-81.

Tong, A, Sainsbury, P & Craig, J 2007 Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care* 19 (6) 349-57.

Tullis, T & Albert, B 2013 *Measuring the user experience - collecting, analyzing, and presenting usability metrics, 2nd Edition* Elsevier: Morgan Kaufmann Publishers, Burlington, MA, USA.

Tóth, G M 2013 The computer-assisted analysis of a medieval commonplace book and diary (MS Zibaldone Quaresimale by Giovanni Rucellai). *Literary and Linguistic Computing* 28 (3) 432-43.

Ultsch, A 1993 Self-Organizing Neural Networks for Visualisation and Classification. in **Opitz, O, Lausen, B & Klar, R** eds *Information and Classification: Concepts, Methods and Applications Proceedings of the 16th Annual Conference of the "Gesellschaft für Klassifikation e.V." University of Dortmund, April 1–3, 1992*. Springer Berlin Heidelberg, Berlin, Heidelberg.

van den Bosch, A, Bogers, T & de Kunder, M 2016 Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics* 1-18.

Wald, A & Wolfowitz, J 1940 On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics* 11 (2) 147-62.

Wallach, H M, Mimno, D & McCallum, A 2009a Rethinking LDA: Why Priors Matter. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*. Vancouver, Canada.

Wallach, H M, Murray, I, Salakhutdinov, R & Mimno, D 2009b Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Canada 1105-12.

Wang, N, Biggs, T W & Skupin, A 2013 Visualizing gridded time series data with self organizing maps: An application to multi-year snow dynamics in the Northern Hemisphere. *Computers, Environment and Urban Systems* 39 107-20.

Wasserman, S & Faust, K 1994 *Social Network Analysis - Methods and Applications* Cambridge University Press, Cambridge, UK.

Weingart, S & Jorgensen, J 2013 Computational analysis of the body in European fairy tales. *Literary and Linguistic Computing* 28 (3) 404-16.

Wharton, C, Rieman, J, Lewis, C & Polson, P 1994 The Cognitive Walkthrough Method: A Practitioner's Guide. in **Nielsen, J & Mack, R L** eds *Usability inspection methods*. John Wiley & Sons, New York, NY, USA.

Wheeler, D C 2014 Geographically Weighted Regression. in **Fischer, M M & Nijkamp, P** eds *Handbook of Regional Science*. Springer Berlin Heidelberg, Berlin, Heidelberg 1435-59.

Wise, J A, Thomas, J J, Pennock, K, Lantrip, D, Pottier, M, Schur, A & Crow, V 1995 Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents. *Proceedings of the 1995 IEEE Symposium on Information Visualization*. Atlanta, GA, USA.

Wong, D 2009 The Modifiable Areal Unit Problem (MAUP). in **Fotheringham, A S & Rogerson, P A** eds *The SAGE handbook of Spatial Analysis*. SAGE Publications Ltd, London, UK.

Wong, Y Y 1992 Rough and ready prototypes: lessons from graphic design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Monterey, CA, USA.

Yang, S Q & Li, L 2016 Chapter 6 - Evolving Digital Library and Library Digitization. *Emerging Technologies for Librarians*. Chandos Publishing 69-102.

Zobel, J & Moffat, A 2006 Inverted files for text search engines. *ACM Computing Surveys (CSUR)* 38 (2).

Appendix

A 203 most frequent toponyms and categories

Rank	Toponym	Category	Frequency	Rank	Toponym	Category	Frequency
1	Zürich	M > 50,000	18340	54	Trogen	M 2,000-9,999	409
2	Bern	M > 50,000	13908	55	Sursee	M 2,000-9,999	397
3	Basel	M > 50,000	12216	56	Moudon	M 2,000-9,999	384
4	Genf	M > 50,000	10386	57	Le Locle	M 10,000-50,000	363
5	Lausanne	M > 50,000	6836	57	Pfäfers	M < 2,000	363
6	St. Gallen	M > 50,000	5204	59	Visp	M 2,000-9,999	358
7	Freiburg	M 10,000-50,000	5069	60	Wädenswil	M 10,000-50,000	357
8	Luzern	M > 50,000	4984	60	Raron	M < 2,000	357
9	Neuenburg	M 10,000-50,000	3548	62	Monthey	M 10,000-50,000	351
10	Solothurn	M 10,000-50,000	3048	63	Baar	M 10,000-50,000	345
11	Chur	M 10,000-50,000	2619	64	Arbon	M 10,000-50,000	327
12	Schaffhausen	M 10,000-50,000	2493	64	Avenches	M 2,000-9,999	327
13	Schwyz	M 10,000-50,000	2315	66	Sankt Urban	CV 100-2,000	319
14	Sitten	M 10,000-50,000	2076	66	Willisau	M 2,000-9,999	319
15	Winterthur	M > 50,000	2026	68	Martigny	M 10,000-50,000	314
16	Lugano	M > 50,000	1939	69	Siders	M 10,000-50,000	304
17	Glarus	M 2,000-9,999	1537	70	Nidau	M 2,000-9,999	296
18	Appenzell	M 2,000-9,999	1337	71	Stein am Rhein	M 2,000-9,999	293
19	Aarau	M 10,000-50,000	1334	72	Rheinau	M < 2,000	292
20	Einsiedeln	M 10,000-50,000	1206	72	Romainmôtier	CV 100-2,000	292
21	Bellinzona	M 10,000-50,000	1011	74	Kreuzlingen	M 10,000-50,000	285
22	Locarno	M 10,000-50,000	992	74	Leuk	M 2,000-9,999	285
23	Pruntrut	M 2,000-9,999	980	76	Wettingen	M 10,000-50,000	277
24	Vevey	M 10,000-50,000	926	77	Sargans	M 2,000-9,999	275
25	Frauenfeld	M 10,000-50,000	854	78	Langenthal	M 10,000-50,000	272
26	Thun	M 10,000-50,000	832	78	Orbe	M 2,000-9,999	272
27	Burgdorf	M 10,000-50,000	756	80	Aigle	M 2,000-9,999	269
28	Herisau	M 10,000-50,000	701	81	Horgen	M 10,000-50,000	258
29	Stans	M 2,000-9,999	692	81	La Neuveville	M 2,000-9,999	258
30	Olten	M 10,000-50,000	678	83	Altstätten	M 10,000-50,000	250
31	La Chaux-de-Fonds	M 10,000-50,000	668	83	Montreux	M 10,000-50,000	250
31	Biel	M > 50,000	668	85	Entlebuch	M 2,000-9,999	245
33	Rapperswil	CV > 2,000	610	86	Näfels	M 2,000-9,999	242
34	Delsberg	M 10,000-50,000	583	86	Aubonne	M 2,000-9,999	242
35	Engelberg	M 2,000-9,999	539	88	Cham	M 10,000-50,000	238
36	Lenzburg	M 2,000-9,999	536	89	Moutier	M 2,000-9,999	237
37	Murten	M 2,000-9,999	518	90	Uznach	M 2,000-9,999	233
37	Brig	CV > 2,000	518	90	Valangin	M < 2,000	233
39	Payerne	M 2,000-9,999	495	92	Maienfeld	M < 2,000	232
40	Sarnen	M 2,000-9,999	484	93	Küsnacht (ZH)	M 10,000-50,000	231
41	Greyerz	M < 2,000	480	94	Ascona	M 2,000-9,999	226
42	Beromünster	M 2,000-9,999	452	95	Weinfelden	M 2,000-9,999	225
43	Morges	M 10,000-50,000	449	96	Wil (SG)	M 10,000-50,000	223
44	Liestal	M 10,000-50,000	443	97	Bischofszell	M 2,000-9,999	222
44	Nyon	M 10,000-50,000	443	98	Aarberg	M 2,000-9,999	221
46	Interlaken	M 2,000-9,999	441	98	Cossonay	M 2,000-9,999	221
47	Disentis	CV > 2,000	439	100	Diessenhofen	M 2,000-9,999	220
48	Mendrisio	M 2,000-9,999	431	101	Laufenburg	M 2,000-9,999	216
49	Zofingen	M 2,000-9,999	429	102	Menzingen	M 2,000-9,999	214
50	Davos	M 10,000-50,000	427	103	Uster	M 10,000-50,000	207
51	Rheinfelden	M 2,000-9,999	423	104	Brugg	M 2,000-9,999	204
51	Altdorf (UR)	M 2,000-9,999	423	105	Wolhusen	M 2,000-9,999	201
53	Rorschach	M 2,000-9,999	413	106	Bex	M 2,000-9,999	199

M = municipality, CV = cities and villages, numbers = population

Rank	Toponym	Category	Frequency	Rank	Toponym	Category	Frequency
107	Rothenburg	M 2,000-9,999	197	156	Balsthal	M 2,000-9,999	135
108	Arlesheim	M 2,000-9,999	195	157	Walenstadt	M 2,000-9,999	134
109	Riehen	M 10,000-50,000	185	158	Sachseln	M 2,000-9,999	132
109	Aarwangen	M 2,000-9,999	185	159	Schönenwerd	M 2,000-9,999	131
109	Arth	M 10,000-50,000	185	159	Bremgarten (AG)	M 2,000-9,999	131
112	Sempach	M 2,000-9,999	183	159	Trub	M < 2,000	131
112	Simplon	M < 2,000	183	162	Klingnau	M 2,000-9,999	130
114	Bülach	M 10,000-50,000	179	162	Fraubrunnen	M < 2,000	130
114	Aarburg	M 2,000-9,999	179	164	Steckborn	M 2,000-9,999	129
114	St. Moritz	M 2,000-9,999	179	164	Fischingen	M < 2,000	129
117	Bellelay	CV 100-2,000	178	166	Gersau	M < 2,000	128
118	Saint-Ursanne	M < 2,000	174	166	Boudry	M 2,000-9,999	128
118	Rhäzüns	M < 2,000	174	168	Frutigen	M 2,000-9,999	125
120	Gais	M 2,000-9,999	173	169	Ernen	M < 2,000	124
120	Zollikon	M 10,000-50,000	173	170	Schleitheim	M < 2,000	122
122	Regensberg	M < 2,000	168	170	Königsfelden	CV < 100	122
123	Ilanz	M 2,000-9,999	167	170	Mollis	M 2,000-9,999	122
124	Hofwil	CV 100-2,000	164	170	Netstal	M 2,000-9,999	122
125	Kerns	M 2,000-9,999	163	170	Biasca	M 2,000-9,999	122
126	Schiers	M 2,000-9,999	162	175	Grüningen	M 2,000-9,999	120
127	Schänis	M 2,000-9,999	161	175	Buochs	M 2,000-9,999	120
128	Münchenbuchsee	M 2,000-9,999	160	177	Signau	M 2,000-9,999	119
128	Lutry	M 2,000-9,999	160	177	La Sarraz	M < 2,000	119
130	Wattwil	M 2,000-9,999	157	179	Knonau	M < 2,000	118
130	Grenchen	M 10,000-50,000	157	179	Unterägeri	M 2,000-9,999	118
132	Herzogenbuchsee	M 2,000-9,999	154	179	Mels	M 2,000-9,999	118
133	Saanen	M 2,000-9,999	153	179	Erstfeld	M 2,000-9,999	118
134	Ruswil	M 2,000-9,999	152	183	Weggis	M 2,000-9,999	117
134	Estavayer-le-Lac	M 2,000-9,999	152	184	Konolfingen	M 2,000-9,999	116
136	Montagny	M < 2,000	152	185	Blenio	M < 2,000	114
137	Ennenda	M 2,000-9,999	151	186	Männedorf	M 2,000-9,999	113
137	Poschiavo	M 2,000-9,999	151	186	Laupen	M 2,000-9,999	113
139	Waldenburg	M < 2,000	150	188	Münchenstein	M 10,000-50,000	112
140	Andelfingen	M < 2,000	149	188	Courtelary	M < 2,000	112
140	Dornach	M 2,000-9,999	149	188	Pully	M 10,000-50,000	112
142	Sissach	M 2,000-9,999	146	191	Thusis	M 2,000-9,999	111
143	Hitzkirch	M 2,000-9,999	145	192	Zuoz	M < 2,000	110
143	Echallens	M 2,000-9,999	145	193	Binningen	M 10,000-50,000	109
145	Lichtensteig	M 2,000-9,999	144	193	Richterswil	M 2,000-9,999	109
146	Köniz	M 10,000-50,000	143	193	Le Landeron	M 2,000-9,999	109
146	Münsingen	M 10,000-50,000	143	196	Muttenz	M 10,000-50,000	108
148	Rheineck	M 2,000-9,999	142	197	Samedan	M 2,000-9,999	106
149	Hochdorf	M 2,000-9,999	139	198	Churwalden	M < 2,000	105
150	Zermatt	M 2,000-9,999	138	199	Thayngen	M 2,000-9,999	104
151	Sumiswald	M 2,000-9,999	137	200	Neunkirch	M < 2,000	100
151	Spiez	M 10,000-50,000	137	201	Langnau im Emmental	M 2,000-9,999	94
153	Stäfa	M 10,000-50,000	136	202	Fleurier	M 2,000-9,999	87
153	Kriens	M 10,000-50,000	136	203	Couvet	M 2,000-9,999	84
153	Silenen	M 2,000-9,999	136				

M = municipality, CV = cities and villages, numbers = population size

Remarks: The category and population data in the table are based on *SwissNames* 2008. The frequency values were assessed after excluding *places of citizenships* in *biographies* and the information about the membership of municipalities to cantons and to districts in *geographical entities* articles.

B Descriptive words of topics in German

Topic	Most descriptive terms
1	römisch gebiet reich kelt inschrift helvetier kaiser provinz alemanne volk name kelte germanisch quelle civitas grab frühmittelalter archäologisch belegen könig
2	politisch partei gründen mitglied bewegung verein verband national organisation gesellschaft konservativ sozial gründung liberal weltkrieg bürgerlich sektion entstehen seit neu
3	bund artikel recht kantonal eidgenössisch staat bundesverfassung politisch staatlich gesetz erst kraft öffentlich bundesgesetz regeln bundesstaat bundesrat verfassung neu schutz
4	wirtschaftlich weltkrieg stark hoch anteil steigen groß markt rund industrie gut entwicklung wachstum bevölkerung land wirtschaft maßnahme sektor krise landwirtschaft
5	bank münze franke snb geld international franken währung gold prägen pfennig milliarde kredit kantonalbanken wechselkurs grossbank börse million nationalbank währungspolitik
6	frau kind mann familie ehe haushalt sozial mutter eltern geburt fürsorge geschlecht person leben arbeit arm vater jung heirat männlich
7	zeitung erscheinen zeitschrift gründen auflage blatt exemplar tageszeitung neu titel verlag tagblatt seit politisch presse übernehmen nachricht organ erst druckerei
8	recht beispiel gericht geistlich weltlich spätmittelalter grundherrschaft herr gut urkunde gebiet abgabe könig lateinisch bischof grundherr adel graf frei römisch
9	johann geschichte gesellschaft historisch jakob heinrich werk begriff deutsch neu johannes friedrich zürcher finden wilhelm hans aufklärung französisch karl rudolf
10	medizin krankheit spital arzt mensch hygiene tod seit bad gesundheitswesen gesundheit anstalt epidemie pest tier kranke bevölkerung tuberkulose gefängnis wissenschaftlich
11	fest spiel brauch sport kleidung eidgenössisch seit feiern tanz anlass hotel tag gast tragen populär fahne tourismus farbe feiertag form
12	armee militärisch eidgenössisch truppe dienst waffe mann general zivil krieg landesverteidigung offizier ausbildung militärdepartement verteidigung weltkrieg soldat regiment infanterie artillerie
13	international bundesrat neutralität jude staat deutsch land weltkrieg europa europäisch flüchtling krieg jüdisch organisation beziehung zusammenarbeit behörde internat ausland konferenz
14	sozial politisch ländlich städtisch bürger wirtschaftlich gesellschaft familie obrigkeit zunft gesellschaftlich neuzeit elite begriff bürgerrecht untertan bürgerlich konflikt unruhe spätmittelalter
15	unternehmen gründen firma million übernehmen seit groß betrieb franken franke erst weltweit produktion gesellschaft aktiengesellschaft beschäftigen ausland umsatz rund weltkrieg
16	gewerkschaft arbeiter sozial arbeitnehmer verband arbeit weltkrieg berufl arbeitgeber streik seit angestellte arbeitszeit sozialversicherung krankenkasse frau arbeitslosigkeit sgb ahv öffentlich
17	bau beispiel städtisch haus entstehen landschaft anlage gebäude raum dienen bauen wohnung boden errichten platz wasserversorgung wasser küche garten ländlich
18	ort eidgenossenschaft eidgenössisch französisch eidgenosse schlacht könig bund frieden vertrag orten tagsatzung herzog krieg militärisch konflikt drei bündnis beide truppe
19	rat verwaltung politisch helvetisch wählen verfassung bundesversammlung regierung bundesrat parlament eidgenössisch aufgabe wahl mitglied beziehungsweise landsgemeinde demokratie kantonal klein ober
20	universität schule gründen ausbildung seit forschung gesellschaft fakultät hochschule eidgenössisch erst akademie eth gründung wissenschaft hoch unterricht wissenschaftlich technisch entstehen
21	landwirtschaftlich bauer landwirtschaft bäuerlich getreide vieh ackerbau viehwirtschaft nutzung weide gebiet milch allmend käse getreidebau genossenschaft boden produktion neuzeit betrieb
22	seit museum international gründen sammlung radio historisch pro film post musik stiftung archiv theater erst national verein bibliothek privat fernsehen
23	kunst künstler architektur werk malerei erhalten künstlerisch architekt schaffen hans stil französisch denkmal san/santo/santa darstellung basler bildhauer so(g)n/sontga maler ausstellung
24	italienisch deutsch literatur sprache rätoromanisch französisch werk text dialekt autor kulturell lied deutschsprachig lateinisch literarisch italiana roman schreiben deutschschweiz svizzera
25	handwerk zunft produktion städtisch markt handel kaufleute betrieb beruf gewerbe herstellung produkt meister heimarbeit handwerkl textilindustrie handwerker geselle arbeitskraft arbeiten
26	bau erst straße eisenbahn verkehr bahn betrieb kohle groß sbb brücke energie technisch entstehen transport schiff elektr automobil strecke linie
27	gewicht masse system ehemalig eigentum metr erbe französisch erben grundstück zgb rente grundbesitz erbrecht circa viertel maß schuldner italienisch pfund
28	erst neu groß beispiel teil führen kommen ende spät zeit bleiben mehr wichtig weit seit früh meist alt bedeutung finden
29	circa mensch kultur bronzezeit neolithikum fund liegen groß archäologisch belegen jahrtausend eiszeit klein eisenzeit nachweisen letzt tier zeit mittler etwa
30	kirche katholisch römisch reformieren kirchlich religiös reformation bischof konzil christlich pfarrer orden bistum geistlich reform konfessionell weltlich papst leben seit

C Insights gained for the spatialized network interface

Erkenntnis	Komp.	Tiefe	Unerw.	Relev.	Tot.
Fischingen und Basel sind trotz geographisch grosser Distanz über das Thema Religion miteinander verbunden im 20. Jahrhundert.	3	4	5	5	17
Im 18. Jahrhundert ist die Position von Zürich im Netzwerk der Schweizer Ortschaften sehr dezentral und Zürich und Bern sind nicht direkt verbunden, was sich zum 19. und vor allem 20. Jahrhundert sehr stark ändert und Zürich ins Zentrum des Netzwerkes der Ortsverbindungen der Schweiz rückt.	5	4	3	5	17
Im 18. Jahrhundert sind das Unterwallis (Monthey) und das Oberwallis (z.B. Ernen) über eine Vogtei des Oberwallis im Unterwallis miteinander verbunden.	2	4	5	5	16
Im 18. Jahrhundert ist die Netzwerkstruktur der Ortsverbindungen der Schweiz linear/schlangenförmig, im 19. Jahrhundert gibt es eine stärkere Verästelung und eine Konzentration auf zentrale Ortschaften, im 20. Jahrhundert konzentrieren sich viele Ortschaften auf 2-3 starke Zentren.	5	4	2	5	16
Lausanne und Genf sind im 19. Jahrhundert nicht direkt mit Zürich verbunden im Vergleich zum 20. Jahrhundert, sondern hängen an Bern.	4	3	4	5	16
Konfessionelle Verbindungen prägen die Netzwerke, die geographische Distanz wird z.T. überspielt (z.B. Stans-Appenzell im 20. Jahrhundert und Einsiedeln-St. Gallen im 19. Jahrhundert trotz geographischer Distanz nahe im Netzwerk zueinander).	3	5	3	5	16
Bei kleinen Ortschaften wie z.B. Regensberg hat die herrschaftliche Komponente (d.h. Landvogtei) einen sehr starken Einfluss auf die Schaffung von Ortsverbindungen durch die Erwähnung in Biographien im 18. Jahrhundert.	3	4	4	4	15
Zürich ist im 18. Jahrhundert sehr stark mit Ortschaften im eigenen Untertanengebiet verbunden, jedoch ansonsten eher dezentral im Netzwerk der Ortsverbindungen der gesamten Schweiz.	2	3	4	5	14
Die sehr zentrale Rolle von Freiburg im Netzwerk des 20. Jahrhunderts ist durch den starken Fokus des HLS auf Biographien zu begründen.	3	4	2	5	14
Küsnacht (ZH) nimmt in den Ortsverbindungen des Kantons Zürich des 19. und 20. Jahrhunderts eine sehr zentrale Rolle ein, was vor allem auf das Thema Bildung (z.B. Lehrerseminar in Küsnacht (ZH)) in Biographien zurückgeführt werden kann.	4	4	3	3	14

Erkenntnis	Komp.	Tiefe	Unerw.	Relev.	Tot.
Hinsichtlich der Nicht-Biographien gibt es eine sehr grosse Streuung an Themen, welche Verbindungen verursachen (z.B. Liturgie) auch beispielsweise, wo kein direkter Kontakt von Personen oder Institutionen auszumachen ist.	4	4	4	2	14
Die Rolle von Solothurn im Netzwerk der Ortsverbindungen wird vom 18. zum 20. Jahrhundert hin weniger zentral, Olten hingegen gewinnt in derselben Zeitspanne an Zentralität.	4	4	2	3–5	14
Im 18. Jahrhundert sind die Innerschweiz und das Tessin über Vogteien der Innerschweiz im Tessin miteinander verbunden. Diese Verbindungen basieren hauptsächlich auf Biographien (z.B. Landvögte oder Heiratsbeziehungen)	2	4	2	5	13
Dass Pruntrut die Residenzstadt des Bischofs von Basel von 1528-1798 war, verursacht die sehr zentrale Position Pruntruts im Netzwerk der Ortsverbindungen des 18. Jahrhunderts.	3	3	2	5	13
Im 19. Jahrhundert basieren mehrere Ortsverbindungen auf der Ebene "Kanton Zürich", welche von Küsnacht (ZH) ausgehen, auf dem Lehrerseminar in Küsnacht (ZH).	3	4	3	3	13
Im 20. Jahrhundert scheint die geographische Nähe immer noch eine grosse Rolle für die Stärke der Ortsverbindungen zu spielen (z.B. starke Verbundenheit der Seegemeinden des Kantons Zürich), was mit der noch immer eingeschränkten Mobilität zusammenhängen kann.	3	3	3	4	13
Freiburg gruppiert im 20. Jahrhundert sehr stark und führt dazu, dass Ortschaften, welche geographisch weiter entfernt von Freiburg liegen, nahe im Netzwerk zu Freiburg dargestellt sind (z.B. Engelberg ist mit Freiburg und nicht mit der Innerschweiz oder Luzern verbunden).	2	3	3	5	13
Im 18. Jahrhundert zeigt sich ein klares katholisches Verbindungsgeflecht Luzern-Solothurn-Pruntrut-Freiburg-Wallis im Netzwerk.	3	2	2	5	12
Freiburg erscheint im 20. Jahrhundert sehr zentral, da es sehr viele Verbindungen aufgrund von Personen, welche in Freiburg studieren, zu anderen Ortschaften gibt.	2	3	2	5	12
Je weiter zurück in der Zeit, desto stärker ist Chur mit kleineren Ortschaften im Netzwerk verbunden und wird insgesamt weniger zentral im Netzwerk der Ortsverbindungen der Schweiz.	5	4	3		12

Erkenntnis	Komp.	Tiefe	Unerw.	Relev.	Tot.
Sehr viele kleine Ortsverbindungen basieren ausschliesslich auf Biographien, für verschiedene Ortschaften sind es Verbindungen über Gymnasien (z.B. Schiers-Zürich und Appenzell-Stans, jeweils im 20. Jahrhundert).	3	4	4	1	12
Freiburg und Luzern sind in allen drei betrachteten Jahrhunderten stark über religiöse Themen miteinander verbunden.	3	3	2	4	12
Bellinzona und Schwyz sind im 18. Jahrhundert aufgrund der "Ennetbirgischen Vogteien" miteinander verbunden.	2	4	2	3	11
Zürich ist im 18. Jahrhundert sehr zentral in einem sternförmigen Netzwerk als Territorialherrschaft zu Vogteien im eigenen Untertanengebiet verbunden.	2	2	2	5	11
Im 19. Jahrhundert sind im Netzwerk politisch-religiöse Verbindungen (z.B. Luzern-Freiburg-Wallis) deutlich erkennbar.	2	2	2	5	11
Religiöse und z.T. auch politische Verbindungen sind im HLS sehr zentral, wirtschaftliche Verbindungen hingegen eher selten.	4	3	4		11
Die Ortschaftsgruppen sind im 18. Jahrhundert sehr regional.	3	3	2	2	10
Insbesondere kleine Ortschaften haben zum Teil ausschliesslich biographische Verbindungen zu anderen Ortschaften, was auch ein Grund dafür ist, dass sie überhaupt in das Netzwerk aufgenommen wurden.	4	3	3		10
Im 18. Jahrhundert hat der Jura eine sehr dominante Rolle im Netzwerk.	3	2	4		9
Im 19. Jahrhundert erscheint Bern als neue Hauptstadt sehr zentral im Netzwerk im Vergleich zum 18. Jahrhundert.	1	1	1	5	8
In der Innerschweiz sind die konfessionellen Verbindungen sehr stark (z.B. Stans-Luzern über Kapuzinerorden in Luzern).	3	3	2		8
Das Tessin erscheint im 19. Jahrhundert als isolierte Ortschaftsgruppe mit sehr wenigen Verbindungen nach aussen.	2	2	3		7

Remark: Some fields do not contain a rating because the historian which rated the relevance is not an expert on some topics and thus left some fields blank. Therefore, further experts would need to be asked in order to obtain ratings for all insights.

Curriculum vitae

André BRUGGMANN
born July 7, 1988
citizen of Degersheim SG

Education

09/2012 – 03/2017 **PhD**, University of Zurich, Department of Geography, Geographic Information Visualization & Analysis

Thesis: *Visualization and Interactive Exploration of Spatio-Temporal and Thematic Information in Digital Text Archives*

supervised by Prof. Dr. Sara Irina Fabrikant
Promotion committee: Prof. Dr. Ross Stuart Purves, Dr. Katja Hürlimann

06/2011 – 07/2012 **Master of Science** in Geography, University of Zurich

Thesis: *Netzwerkvisualisierung der Ostschweiz – Die Raumgliederung der Schweiz mit Wikipedia neu formuliert*

supervised by Dr. Marco Michele Salvini, Prof. Dr. Sara Irina Fabrikant

Minor: *Environmental Sciences*

09/2007 – 05/2011 **Bachelor of Science** in Geography, University of Zurich

Thesis: *Explorative Geovisualisierung im Kontext der "Cities and the Creative Class"*

supervised by Dr. Marco Michele Salvini

Minor: *Economics and Business Administration*

08/2003 – 06/2007 **Matura**, Kantonsschule Romanshorn

Major: *Economics & Law*

Publications

Bruggmann, A & Fabrikant, S I 2016a How does GIScience support spatio-temporal information search in the humanities? *Spatial Cognition & Computation*.

Bruggmann, A & Fabrikant, S I 2016b Die Geschichte der Schweiz aus einer geographischen Perspektive betrachtet. *Geomatik Schweiz* 4/2016 108-10.

Bruggmann, A & Fabrikant, S I 2014a How to visualize the geography of Swiss history. in **Huerta, J, Schade, S & Granell, C** eds Connecting a Digital Europe through Location and Place. International Conference on Geographic Information Science, *AGILE 2014*, Castellón, Spain, ISBN: 978-90-816960-4-3.

Bruggmann, A & Fabrikant, S I 2014b Spatializing a Digital Text Archive about History. Geographic Information Science (GIScience 2014) pre-conference workshop on Geographic Information Observatories 2014, Vienna, Austria. *CEUR Workshop Proceedings* 1273 15-22.

Bruggmann, A & Fabrikant, S I 2014c Spatializing time in a history text corpus. in **Stewart, K, Pebesma, E, Navratil, G, Fogliaroni, P & Duckham, M** eds *Proceedings of the 8th International Conference on Geographic Information Science (GIScience 2014)*. Vienna, Austria 183-86.

Bruggmann, A, Salvini, M M & Fabrikant, S I 2013a Agglomerationen mit nutzergenerierten Inhalten neu definiert – Visualisierung der Nordostschweiz mithilfe von Wikipedia. *disp – The Planning Review* 49(4) 37-45.

Bruggmann, A, Salvini, M M & Fabrikant, S I 2013b Cartograms of self-organizing maps to explore user-generated content. *Proceedings of the 26th International Cartographic Conference (ICC)*. Dresden, Germany.